心理学报 2025, Vol. 57, No. 11, 2043–2059 https://doi.org/10.3724/SP.J.1041.2025.2043
Acta Psychologica Sinica https://cstr.cn/32110.14.2025.2043

人类对大语言模型的热情和能力感知*

武月婷[1] 王 博[2,3] 包寒吴霜[4] 李若男[1]
吴 怡[1] 王嘉琪[1] 程 诚[3] 杨 丽[3,5]

([1]天津大学教育学院; [2]天津大学智能与计算学部; [3]天津大学应用心理研究所, 天津 300354)
([4]华东师范大学心理与认知科学学院, 上海 200062) ([5]天津市自杀心理与行为研究实验室, 天津 300354)

摘 要 随着大语言模型(Large Language Models, LLMs)能力的提升及其广泛应用, 社会正逐步从传统的人际交互转向融合人际交互、人机交互和机机交互的多层次互动结构。在人类与 LLMs 交互日益深入的背景下, 研究人类如何感知 LLMs 成为了重要议题。本研究通过三项研究系统考察人类对 LLMs 的感知模式。研究 1 发现, 与对人类的感知一致, 人类主要通过热情和能力两个维度感知 LLMs。然而, 在一般情境下, 不同于对人类感知中的热情优先, 人类在对 LLMs 的感知中能力优先。研究 2 探讨了热情和能力在不同态度预测中的优先效应, 结果表明, 热情与能力均能正向预测人类对 LLMs 的持续使用意愿和喜爱度, 其中能力对持续使用意愿的预测效力更高, 而热情对喜爱度的预测效力更高。研究 3 进一步探索了人类对 LLMs 与对他人的感知差异, 结果显示, 人类对 LLMs 的热情评价与人类无显著差异, 但对 LLMs 的能力评价显著高于人类。本研究为理解人类对 LLMs 的感知提供了理论基础, 并为人工智能的设计优化及人机协作机制的研究提供了新的视角。

关键词 大语言模型, 社会认知, 热情, 能力
分类号 B849: C91

1 引言

随着人工智能和自然语言处理技术(Natural Language Processing, NLP)的快速发展, 大语言模型(Large Language Models, LLMs)的研发与应用进展显著(Brown et al., 2020; Ouyang et al., 2022)。LLMs 的发展正在推动通用人工智能的实现(Bubeck et al., 2023)。LLMs 开始表现出类似人类的心理特征, 如在心理理论(Theory of Mind, ToM)能力方面已接近 6 岁儿童(Kosinski, 2024), 甚至在某些任务中的表现超越人类水平(Strachan et al., 2024)、表现出类似人类的人格特质(Huang et al., 2023; Jiang et al., 2023)、具有与人类相似的根据情感线索调节反应的能力(Zhao et al., 2024)。

随着 LLMs 逐步融入日常生活, 人们的交互对象也从传统的自我、他人和群体扩展到 LLMs。这将深刻影响社会互动形态, 并带来新的挑战。在此背景下, 人类如何感知 LLMs 成为一个亟待回答的问题, 即人类是否会像感知他人一样, 采用类似社会认知内容的理论框架来感知 LLMs?这一问题具有重要意义, 不仅有助于理解人类对智能体的社会认知方式, 还可为人工智能的设计优化和人机协作机制的研究提供重要的理论启示。

1.1 人类对 LLMs 的感知维度

在社会生活中, 人们不断对他人、群体以及自我进行感知并做出评价, 这些评价直接影响我们的态度和行为, 并指导我们做出决策。因此, 探讨人们如何感知自我与他人是理解人类社会互动的关

收稿日期: 2024-02-06
* 国家自然科学基金面上项目(项目编号：62376188), 国家社会科学基金一般项目(项目编号：21BSH017, 22BSH163), 2024 年第一批天津市制造业高质量发展专项建设智能化数字化应用场景—自主智能算力的通用大模型关键技术研究及产业化应用示范项目(项目编号 24ZGNGX00020)支持。
通信作者: 王博, E-mail: bo_wang@tju.edu.cn; 杨丽, E-mail: yangli@tju.edu.cn

键起点。既有理论研究普遍认为, 人际社会评价主要围绕两个维度, 这一框架被称为社会认知内容的“大二”模型(Big Two), 并已在多项跨文化研究中得到验证(Abele et al., 2016; Bai et al., 2020; Cuddy et al., 2009)。其中具有代表性的理论包括双视角模型(Dual Perspective Model, DPM)和刻板印象内容模型(Stereotype Content Model, SCM)。双视角模型认为, 亲和性(Communion)和能动性(Agency)是感知自我与他人的核心维度(Abele &Wojciszke, 2007, 2014, 2018)。其中, 亲和性主要涉及友好、关怀等特质, 而能动性则涉及能力、独立性和主观能动性等特质。刻板印象内容模型则认为, 人们感知群体时主要关注热情(Warmth)和能力(Competence)(Fiske et al., 2002, 2007)。热情指一个群体是否被认为友好和善良, 能力则衡量该群体是否具备完成任务或展现专业素质的水平。尽管双视角模型与刻板印象内容模型在研究对象及维度内涵上存在一定差异, 但近年来, Abele等人(2021)通过“对抗——合作”的方式, 提出了整合框架, 将社会认知内容的两个核心维度明确为纵向维度与横向维度。其中纵向维度也被称为能力/能动性, 横向维度也被称为热情/亲和性。我国学者佐斌等人(2015)在社会认知内容的“大二”模型的综述中指出, “考虑到亲和性和能动性容易与人格心理学以及社会心理学其他领域的概念混淆”, 在研究中使用热情和能力进行命名。本研究沿用了这一命名原则, 使用“热情”与“能力”的翻译。

随着人工智能的广泛应用, 人工智能体逐步融入人类社会, 成为社会交际过程中的重要组成部分。因此, 了解人们如何感知人工智能, 是理解和发展良好人机交互的关键。McKee 等人(2023)验证了热情和能力是人类对传统人工智能体(如智能语音助手、推荐系统、游戏系统、自动驾驶等)的主要感知维度。然而, 目前尚未有研究在 LLMs 领域验证该理论适用性。尽管 LLMs 仍由算法驱动, 但 LLMs 的语言生成能力模糊了人机边界, 在形式上表现出更为类人的行为。同时, 伦理规范和道德属性日益成为 LLMs 研发重要议题, 这促使人们在评估 LLMs 时可能更加关注其道德属性。例如, 人们可能会质疑 LLMs 是否能够正确理解并遵循社会伦理, 而这在“大二”模型中往往被归入热情维度(道德通常被视为热情维度的切面)。此外, LLMs 的发展引发了关于其生成能力的广泛讨论, 进一步突出了其在能力维度上的表现。这些提示, 人类对 LLMs 的感知可能延续“大二”模型的二维结构。因此, 本研究提出以下假设:

H1: 热情和能力是人类对 LLMs 感知的主要维度。

1.2 一般情境下人类对 LLMs 的感知维度中的优先效应

在社会认知内容的研究中, 热情和能力孰轻孰重是一个重要问题, 被称为优先效应。研究表明, 在一般的人际感知情境中, 热情通常优先于能力, 即个体在评价他人时, 往往首先关注对方的热情, 而后才是对方的能力, 同时热情评价会引起个体更多的态度和行为反应(Abele &Wojciszke, 2007; Fiske et al., 2007; Fiske, 2018)。双视角模型通过利益最大化原则对此进行解释(Abele &Wojciszke, 2007)。具体而言, 在人际互动中, 热情为他人有利特质, 而能力为自我有利特质。为实现自身利益, 个体作为观察者时, 会更加关注他人的热情表现, 而非能力表现。

然而, 在人类对 LLMs 的感知中, 热情与能力对谁有利, 与在人类对人类的感知中的划分并不完全一致。由于 LLMs 被假定不存在“自我”, 所以在 LLMs 的使用者(即人类)看来, LLMs 的热情和能力都是服务于人类(他人)而非其自身的, 所以均属于他人有利特质。进一步, 相较于热情, LLMs 的能力通常具有更明显的回馈性和实用性, 例如提供答案、完成任务、辅助决策等。这种高效的回报性能够更直接地帮助用户实现目标, 因而 LLMs 的能力更能够使人类实现利益最大化。因此, 我们提出以下假设:

H2: 在人类对 LLMs 的感知维度中, 能力较之热情具有优先效应。

1.3 预测人类对 LLMs 的态度时感知维度的优先效应

正如 Fiske (1992)指出“知觉是为了行动”, 对社会认知内容的研究最终指向其如何引发认知、情感及行为反应。群际刻板印象−情绪−行为倾向系统模型(Behaviors from Intergroup Affect and Stereotypes Map, BIAS Map)进一步探讨了感知与情绪及行为反应之间的关系(Cuddy et al., 2007; Fiske et al., 2007)。BIAS Map 指出, 对热情的感知影响人们的主动性行为倾向, 而对能力的感知则影响被动性行为倾向。同时, BIAS Map 根据热情与能力两个维度的不同组合, 将社会知觉对象划分为 4 类群体, 并发现人们对不同类型群体会产生相应的

情绪反应。例如, 对高热情且高能力的群体通常产生钦佩情绪, 而对高能力但低热情的群体则产生嫉妒情绪。而针对传统人工智能体的研究亦表明, 热情与能力维度能够预测人类对它们的态度。机器人社会属性量表(Robotic Social Attributes Scale, RoSAS)提出热情、能力感和不适三因子描述人们对机器人的社会属性的感知(Carpinella et al., 2017)。Scheunemann 等人(2020)在此基础上进一步确认, 热情和能力是预测人类与机器人互动偏好的最佳因子。

热情和能力两个维度对态度的预测, 也体现了优先效应的内涵。优先效应不仅优先感知某一维度, 也指某一维度对个体行为和态度反应具有更强影响力(Fiske et al., 2007)。因此, 本文提出以下两个问题, 一是人类对 LLMs 的热情与能力评价能否有效预测人类对 LLMs 的态度倾向? 二是在不同态度预测中, 维度的优先性是否会发生变化? 为了回答这些问题, 本研究选取持续使用意愿与喜爱度作为两个不同的评估目标。这两个评估目标(或称为态度指标)是评估人机交互结果的典型指标。

首先, 持续使用意愿作为一种功能型认知评估指标, 能够反映人类对 LLMs 完成任务能力的认可程度。持续使用意愿受到多个核心因素的影响, 其中包括绩效期望、努力期望、社会影响以及便利条件(Venkatesh et al., 2003)。这些核心因素反映了人类对 LLMs 的功能评估, 尤其是对其任务完成效能与易用性的认知。在这一框架下, 人类对 LLMs 的能力评价直接影响其持续使用意愿。其次, 喜爱度作为一种情感型认知评估指标, 能够反映人类与 LLMs 之间的情感联结程度。与持续使用意愿不同, 喜爱度更多与幽默(Mirnig et al., 2017; Tae & Lee, 2020)和拟人化(Mohammad & Nishida, 2015; Mori, 1970)有关。这些因素与热情特质类似, 且热情高的 LLMs 通常能够带给人类更愉悦的交互体验, 从而引发更强的情感联结。因此, 热情评价在喜爱度形成中起到了更重要的作用。综上, 我们提出以下假设:

H3a: 人类对 LLMs 的热情和能力评分均可以正向预测人类对 LLMs 的持续使用意愿, 且能力评分的预测效力更高。

H3b: 人类对 LLMs 的热情和能力评分均可以正向预测人类对 LLMs 的喜爱度, 且热情评分的预测效力更高。

1.4　人类对他人和对 LLMs 的热情和能力评分的比较

上述研究聚焦于探索对 LLMs 感知的模式, 但是目前尚不清楚在人类感知中, LLMs 在热情与能力两个维度上表现如何。虽然 LLMs 在类人性上取得了长足进展, 但人类复杂的认知、情感与行为特征尚难以被完全模拟。具体表现为, 当前人类对 LLMs 在情感互动上的评价呈现矛盾性特征: 尽管现有研究表明 LLMs 能够生成符合社会规范的共情反应(dos Santos, 2023), 但人类用户常表达 LLMs 的表现程序化且缺乏情感等评价。因此, 尽管 LLMs 能够生成符合社会规范的情感表达, 但由于其缺乏情感深度, 人类可能会倾向于认为在热情维度上, 相较于他人, LLMs 的表现更差。与人类用户在热情维度上对 LLMs 的评价不同, LLMs 引发的巨大关注恰源于人们对其能力提升的感知。相关研究表明, LLMs 在逻辑推理(Bubeck et al., 2023)和抽象能力(Webb et al., 2023)等多个领域已超越了人类一般水平。因此, 我们提出以下假设:

H4a: 人类对人类和 LLMs 的热情评分存在差异, 对 LLMs 的热情评分更低。

H4b: 人类对人类和 LLMs 的能力评分存在差异, 对 LLMs 的能力评分更高。

1.5　研究概述

本研究旨在构建人类对 LLMs 感知的理论框架, 系统分析人类对 LLMs 的感知维度, 并进一步探讨其在一般情境下的优先效应及该效应在态度预测中是否发生变化。同时, 本研究比较了人类对 LLMs 与对人类的感知差异。本研究填补了人类对 LLMs 感知研究的空白, 并为人工智能的设计与优化提供了新的理论视角。

2　研究 1: 人类感知 LLMs 的维度及一般情境下的优先效应

2.1　研究 1a: 基于自由反应任务分析人类对 LLMs 的感知维度及一般情境下感知维度的优先效应

2.1.1　研究目的

本研究旨在通过自由反应任务探索人类感知 LLMs 的主要维度及维度的优先性, 以验证 H1 和 H2, 即验证热情和能力是人们对 LLMs 感知的主要维度, 且能力较之热情具有优先效应。

2.1.2　方法

(1) 被试

通过某问卷调查平台在线招募被试 215 名, 排

除其中 8 名完全没有 LLMs 使用经验和未按指令进行反应的被试, 最终有效被试人数为 207 名, 其中男性 124 名(59.90%), 女性 83 名(40.10%)。所有被试均自愿参与研究。

为了确保正式研究的样本量能够达到适当的统计检验力, 本研究基于预实验数据(31 名被试)进行了检验力分析。由于我们计划使用广义线性混合模型(Generalized Linear Mixed Model, GLMM)来检验研究假设, 因此采用 R 语言 simr (Green & MacLeod, 2016)进行基于模拟的检验力分析(双尾检验, $\alpha = 0.05$; 进行 1000 次蒙特卡罗模拟)。结果显示, 当样本量为 31 时, 检出热情和能力覆盖率的非标准化回归系数为−0.50 的检验力分别为 88.70% 和 89.80%; 检出热情和能力覆盖率的非标准化回归系数为−1.00 时, 检验力均达到 100%。因此, 正式研究中的样本量可以满足统计检验力的要求。

(2)工具

本研究采用 SADCAT (Semi-Automated Dictionary Creation for Analyzing Text)进行分析。SADCAT 是由 Nicolas 等人(2021)开发的专门用于研究刻板印象内容的半自动化英文心理词典。该词典基于刻板印象内容模型中的热情和能力两维度构建, 并使用 WordNet[1]进一步拓展, 最终形成了社会性、道德、才智、自信、地位、外貌等 16 个印象维度的子词典。其中, 社会性和道德两个子维度构成热情维度, 才智和自信两个子维度构成能力维度。SADCAT 可能受到种子词[2]、WordNet 分布以及语义广泛程度(Semantic Generality)差异的影响, 其子词典长度从 7 个到 2402 个单词不等, 分布并不平衡, 但对刻板印象的覆盖率在 82%左右。

(3)操作性定义

目前, 优先效应通常通过处理速度、主观权重以及务实性诊断三种方式验证(Abele et al., 2021)。其中, 处理速度是指认知过程中被优先处理和识别的维度具有优先效应; 主观权重反映感知者对某一维度赋予更高重要性的倾向; 务实性诊断则指在不同个体的评价中, 评价一致性较高的维度具有优先效应。本研究采用处理速度作为优先效应的验证方式, 以考察被试在描述 LLMs 时对不同维度词汇的选择倾向。具体而言, 优先效应在本研究中被定义为: 在被试描述 LLMs 的词汇中, 覆盖率更高的感知维度视为在认知过程中被更迅速注意和处理的维度, 即具有优先效应的维度。

(4)程序

首先, 我们通过问卷要求被试使用至少三个中文词汇描述他们对 LLMs 的印象, 且未限制词汇类型。最终, 共收集到来自 207 名被试的 612 个描述词汇(包括重复词汇)。尽管部分被试仅使用一个或两个词汇进行描述, 但这些词汇仍被保留在数据中。

随后, 我们对被收集到的词汇文本进行进一步处理。具体步骤如下: (1)翻译: 使用翻译工具将中文词汇翻译为英文; (2)校对: 由英语专业人员对翻译结果进行审核和校对, 以确保翻译的准确性; (3)词汇预处理: 将词汇转为小写、去除标点符号、词形还原[3]、去除停止词[4]等; (4)维度识别: 将所有处理后的词汇与 SADCAT 进行匹配, 统计其分别属于热情、能力、社会性、道德、才智、自信等维度的词汇数量; (5)维度覆盖率计算: 将每个维度对应的词汇数量除以该被试反应文本中的总描述词汇数, 得到各维度的覆盖率, 以衡量不同感知维度的占比。

接着, 我们利用维度覆盖率进行广义线性混合模型分析。广义线性混合模型适用于本研究的重复测量数据和非正态分布情况, 能够有效处理数据的非独立性。我们利用 R 包 lme4 进行广义线性混合模型分析, 将覆盖率作为因变量, 维度(如热情、能力等)作为固定效应, 被试作为随机效应, 并将描述词汇数量作为权重, 构建广义线性混合模型。同时, 考虑到本研究中性别比例不平衡, 我们将性别作为协变量纳入模型进行控制分析, 结果显示性别变量的效应不显著。因此, 我们去除性别变量后按照上述方法重新进行模型拟合。

接着, 通过广义线性混合模型, 我们计算出每个维度的估计边际均值, 即在控制了其他变量的情况下, 评估每个维度在被试对 LLMs 的感知中的平均覆盖率, 提供更加准确的效应估计。最后, 为了进一步探讨不同维度对被试在 LLMs 感知中的覆盖

[1] WordNet 是一个在线英文词汇数据库。它对单词按其意义进行分组, 并记录词汇之间的各种语义关系, 如同义词、反义词、上下位词等。

[2]种子词是用于启动词典构建过程的一组基础词汇。

[3]词形还原是指将单词还原为它们的基本形式或词元, 以便在文本分析中保持一致性和避免词汇语义之外的多样性。例如, 将动词的各种时态和形态还原为其原始词元, 如将“running”还原为“run”。本研究使用 R 包 textstem 包进行词形还原。

[4]停止词是指在各类文档中频繁出现但不携带实际信息的高频词汇, 例如“a”、“and”、“of”等。本文使用 R 包 tm 去除这些停止词。

率的相对影响, 我们对热情和能力两个维度的覆盖率差异进行了比较。

2.1.3 结果

我们构建了两个广义线性混合模型：二维模型(固定效应为热情和能力)与子维度模型(固定效应为社会性、道德、自信、地位、才智、信念、外貌等 16 个子维度, 且由于所有被试在家庭维度的覆盖率均为 0, 因此在模型拟合时排除了家庭维度, 最终保留 15 个子维度)。分析结果表明, 两个广义线性混合模型中的固定效应均显著, 具体结果见表 1。子维度的广义线性混合模型 AIC 为 2807.34, BIC 为 2904.00, 均大于二维模型(热情和能力)的广义线性混合模型的 AIC 值 981.89 和 BIC 值 993.97。AIC 和 BIC 是将模型拟合优度和复杂度进行平衡的指标, 较低的 AIC 和 BIC 值表示模型能够更好地拟合数据, 同时模型的复杂度不至于过高。因此, 热情和能力的二维广义线性混合模型的拟合效果更好。

其次, 进一步计算了热情、能力及其他维度的估计平均值, 并进行了指数化处理, 以获取不同维度的实际平均覆盖率。结果显示, 热情和能力维度的覆盖率最高。在子维度层面, 热情维度中的道德切面和能力维度中的才智切面表现出较高的覆盖率。相较之下, 热情维度中的社会性切面和能力维度中的自信切面未表现出显著的较高覆盖率。具体结果参见表 2 和图 1。

最后, 我们比较了能力与热情两个维度的覆盖率差异, 结果表明, 与热情相关的内容明显少于与能力相关的内容(odd ratio = 2.88, z = 9.51, 95% CI [2.32, 3.59], p < 0.001), 即能力的覆盖率更高。

2.1.4 讨论

研究 1a 采用半自动化工具, 将被试反馈的词汇通过现有的词汇分类词典, 自动对应到相关维度进行分析。结果支持假设 H1, 即人类对 LLMs 的主要感知维度与对他人的主要感知维度一致, 均为热情和能力。

然而, 与人类对他人的感知不同, 本研究支持 H2, 人类对 LLMs 的感知中能力优先, 而非热情优先。表面上看, 这种能力优先效应似乎与“大二”模型中的热情优先效应相对立, 但实际上, 二者均遵循利益最大化原则。在人类对他人的感知中, 热情和能力分别是他人有利和自我有利特质, 因此在感知他人时, 为实现自己利益的最大化, 优先感知他人的热情(Abele &Wojciszke, 2007)。但是, 在人类对 LLMs 的感知中, 热情与能力均被视为他人有利特质, 且由于 LLMs 的能力特质具有更高的直接回馈性, 因而对 LLMs 能力的关注可以实现自我利益最大化。

表 1　二维模型以及子维度模型的固定效应汇总

维度	*b*	*SE*	*z*	*p*	95% CI
能力	−0.98	0.06	−15.11	<0.001	[−1.11, −0.85]
热情	−2.04	0.09	−22.53	<0.001	[−2.22, −1.86]
才智(Ability)	−1.14	0.07	−16.87	<0.001	[−1.27, −1.01]
自信(Assertiveness)	−3.14	0.15	−21.69	<0.001	[−3.42, −2.85]
道德(Morality)	−2.22	0.10	−22.84	<0.001	[−2.41, −2.03]
社会性(Sociability)	−3.63	0.18	−19.94	<0.001	[−3.99, −3.27]
地位(Status)	−2.83	0.13	−22.47	<0.001	[−3.07, −2.58]
外貌(Appearance)	−3.60	0.18	−20.06	<0.001	[−3.95, −3.25]
健康(Health)	−3.66	0.19	−19.82	<0.001	[−4.02, −3.30]
职业(Occupation)	−3.50	0.17	−20.42	<0.001	[−3.84, −3.17]
信念(Beliefs)	−4.03	0.22	−18.29	<0.001	[−4.46, −3.59]
异常行为(Deviance)	−4.18	0.24	−17.61	<0.001	[−4.65, −3.72]
情绪(Emotion)	−3.98	0.22	−18.48	<0.001	[−4.40, −3.56]
其他(Other)	−3.98	0.22	−18.49	<0.001	[−4.40, −3.56]
美貌(Beauty)	−5.13	0.38	−13.57	<0.001	[−5.87, −4.39]
地理(Geography)	−6.39	0.71	−9.06	<0.001	[−7.77, −5.01]
社会群体(Social groups)	−6.40	0.71	−9.03	<0.001	[−7.78, −5.01]

注：才智和自信是能力的子维度, 道德和社会性是热情的子维度, 下表同。

表 2 被试描述词汇在 SADCAT 中各印象维度上的覆盖率

维度	估计边际均值	95% CI 下限	95% CI 上限
能力(Competence)	0.38	0.33	0.43
热情(Warmth)	0.13	0.11	0.16
才智(Ability)	0.32	0.28	0.37
自信(Assertiveness)	0.04	0.03	0.06
道德(Morality)	0.11	0.09	0.13
社会性(Sociability)	0.03	0.02	0.04
地位(Status)	0.06	0.05	0.08
职业(Occupation)	0.03	0.02	0.04
外貌(Appearance)	0.03	0.02	0.04
健康(Health)	0.03	0.02	0.04
其他(Other)	0.02	0.01	0.03
情绪(Emotion)	0.02	0.01	0.03
信念(Beliefs)	0.02	0.01	0.03
异常行为(Deviance)	0.02	0.01	0.02
美貌(Beauty)	0.01	0.00	0.01
地理(Geography)	0.00	0.00	0.01
社会群体(Social groups)	0.00	0.00	0.01

此外, 另一种可能的解释是, 在人类的感知中, LLMs 更类似于与自身具有高结果依赖性的群体。人际感知的研究表明, 人际关系中的结果依赖性增强了他人对个人目标实现的相关性, 从而提高了能力维度在整体印象中的重要性(Abele & Brack, 2013)。结果依赖性是指个体在多大程度上受到他人行为的影响, 以及他人能否达成其目标对自己的重要程度。例如, 相较于普通同学, 对朋友的印象中能力维度更重要(Abele &Wojciszke, 2007); 相较于关系较远的人, 对亲密朋友的感知中能力更重要(Wojciszke& Abele, 2008); 在感知存在依赖关系的上级时, 能力维度同样变得更重要(Wojciszke& Abele, 2008)。现阶段, 人类使用 LLMs 主要是为了完成任务, 例如询问意见、信息检索、解决专业问题等(Wang et al., 2024), 这种任务导向的关系使得人类对 LLMs 的结果依赖程度较高, 因而对其感知中能力优先。

尽管本研究支持了假设 H1 和 H2, 但使用的词汇分类工具 SADCAT 可能对研究结果产生干扰。SADCAT 是围绕对人类的感知开发的, 部分词汇在应用于描述 LLMs 语境中时, 其类别归属可能存在偏差。此外, 该词典当前为英文版, 语言与文化差异可能对结果产生干扰。因此, 在研究 1b 中, 我们将采用传统的特质词评定任务进行验证, 以确保结果的稳健性。

2.2 研究 1b：基于特质词评定任务分析人类对 LLMs 的感知维度

2.2.1 研究目的

为避免受到 SADCAT 工具的影响, 本研究通过特质词评定任务直接收集人类对 LLMs 的特质词的判断数据, 以检验 H1, 即验证热情和能力是人们对 LLMs 感知的主要维度。

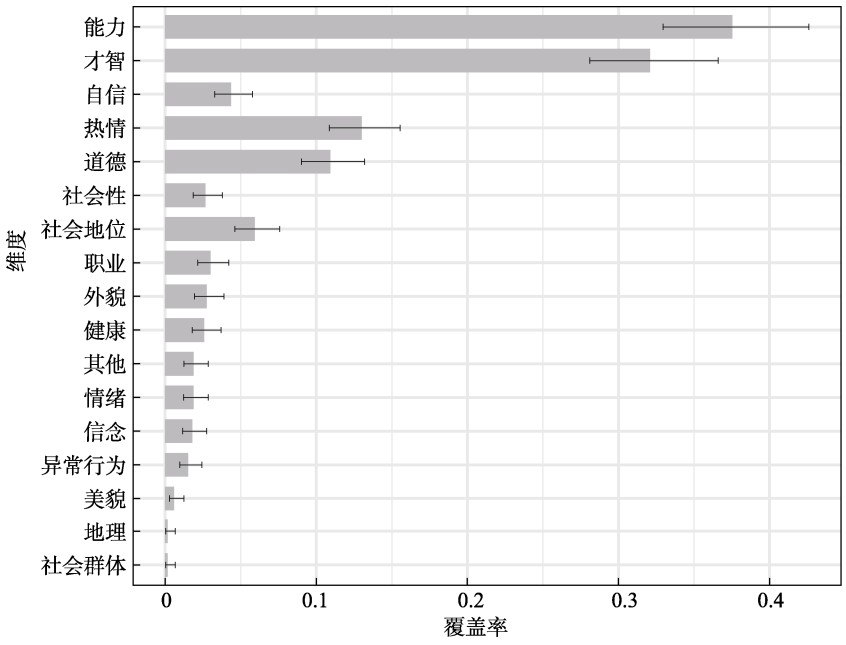

图 1 被试描述词汇在 SADCAT 中各印象维度上的覆盖率

2.2.2 方法

(1)被试

本研究通过某问卷调查平台在线招募具有 LLMs 使用经验的被试。基于统计效能考虑, 按照题量与被试人数比例 1∶5 的原则招募被试, 预估至少需 245 名被试。最终共招募了 300 名被试, 排除 81 名完全没有 LLMs 使用经验和未按指令反应的被试, 最终有效被试人数为 219 名, 其中男性 134 名(61.19%), 女性 85 名(38.81%)。所有被试均自愿参与研究。根据 MacCallum 等人(1999)的研究, 在因子负荷较高且多数共同度良好的情况下, 样本量至少应为 100 到 200; 同时, Comrey 和 Lee (1992)指出, 因子分析一般样本量在 200 左右为适中。因此, 我们研究中的被试人数在可接受范围内。

(2)程序

基于研究 1a 中被试使用的印象词汇, 我们编制了 LLMs 评价词汇池, 具体步骤及操作原因如下。(1)保留形容词, 删除名词、动词等：确保评价集中于描述 LLMs 特征的词汇, 避免名词和动词等与评价性质无关的词语干扰; (2)删除形容词前的修饰副词(例如“有点”、“不太”)：副词通常用表示程度或强度, 在后续评价中被试需选择适用程度, 因此在此步骤中删除副词, 便于后续处理; (3)删除被试行为描述(例如“爱用的”)和效价(例如“好”)类型的词汇：这些词汇往往反映了个体的行为或情感反应, 而非 LLMs 本身的特征, 以提高评价的客观性和描述的准确性; (4)合并反义词：如果保留的词汇中有其反义词, 则将其转换为正效价词汇, 如果没有, 则保留负效价词汇; (5)合并近义词：对保留词汇中存在的近义词进行合并, 确保词汇精简和明确。

经过以上筛选, 最终保留了 29 个词汇。为避免遗漏与 LLMs 特征相关的词汇, 考虑到近期研究发现, 大五人格适用于描述 LLMs 的人格特征(Huang et al., 2023; Jiang et al., 2023)。因此, 我们纳入了中文形容词大五人格简版量表中的 20 个特质词。中文形容词大五人格简版量表包含 20 个条目, 每个条目中有两个对立的形容词(罗杰, 戴晓阳, 2018)。随机选取每个条目中的一个词汇, 将其加入到 LLMs 评价词汇池中。最终, LLMs 评价词汇池包含了 49 个词汇。

接着, 令被试阅读特质词评定任务的指导语, 根据个人经验和理解, 按 5 分制评价词汇池中的每个词汇对 LLMs 的适用程度, 其中 1 表示“完全不适用”, 5 表示“完全适用”。

2.2.3 结果

Bartlett 球形度检验(χ^2 = 7926.14, df = 1176, p < 0.001)和 KMO 检验(KMO = 0.93)表明词汇可能共享潜在因子, 适合进行探索性因素分析。根据 Costello 和 Osborne (2005)的建议, 在 SPSS 中使用极大似然法和最大方差法对 49 个词汇进行探索性因子分析(Exploratory Factor Analysis, EFA)。结果显示, 有 8 个因子的特征根大于 1, 这 8 个因素的解释方差分别为：34.97%、15.20%、4.22%、3.08%、2.76%、2.53%、2.36%、2.13%; 8 个因素累计解释方差为 67.24%。根据碎石图标准(Cattell, 1966), 最终提取了两个因素, 见图 2。

由于我们的研究目的是探索人类感知 LLMs 的主要维度, 我们依据心理测量学理论, 按照下述方法进一步对词汇和因子进行删减使因子结构更加简洁和清晰, 直至完全满足条件：(1)删除项目在所

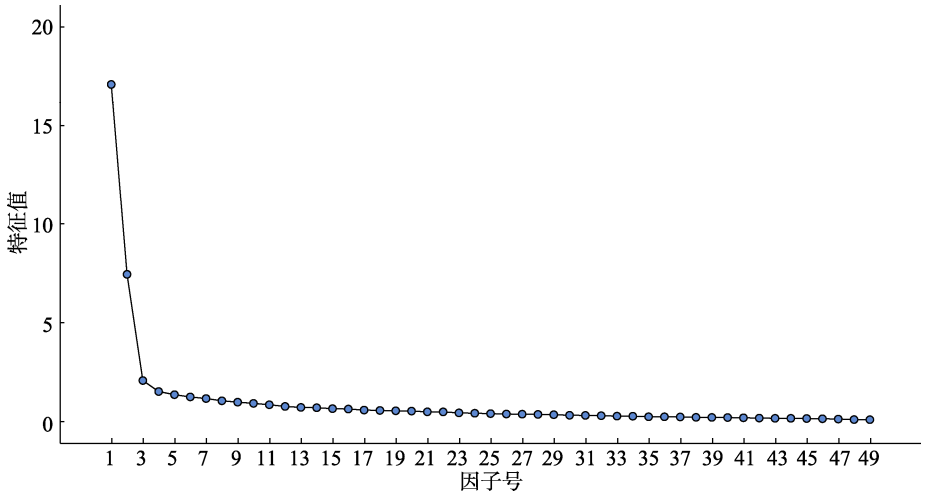

图 2　探索性因子分析碎石图

有因子上负荷都很低(负荷值小于 0.50)的词汇; (2)删除在2个及以上的因子上负荷值高于0.50的词汇; (3)删除少于 3 个及以下词汇的因子; (4)删除共同度小于 0.40 的词汇。基于上述标准, 最终保留 36 个词汇, 累计解释变异量为 53.04%。考虑到本研究中性别比例不平衡, 我们分别对男性样本和女性样本重复上述标准进行探索性因子分析。最终探索性因子分析结果见表 3。

表 3 探索性因子分析结果

词汇	全部(N = 219)			男性(N = 134)			女性(N = 85)		
	因子 1	因子 2	共同度	因子 1	因子 2	共同度	因子 1	因子 2	共同度
智能的	0.83		0.70	0.80		0.65	0.87		0.82
实用的	0.81		0.70	0.80		0.75	0.71		0.66
耐心的	0.78		0.70	0.72		0.76	0.71		0.56
高级的	0.78		0.64	0.73		0.53	0.79		0.75
快速的	0.77		0.75	0.78		0.62			
丰富的	0.77		0.70	0.69		0.49	0.80		0.71
信息多的	0.76		0.60	0.69		0.52	0.78		0.84
易用的	−0.76		0.66	−0.76		0.61			
有效率的	0.76		0.66	0.82		0.68			
全面的	0.76		0.63	0.73		0.55	0.79		0.67
好用的	0.75		0.65	0.78		0.61	0.75		0.62
博学的	0.74		0.60	0.77		0.60	0.71		0.57
有条不紊的	0.72		0.59	0.66		0.54	0.76		0.66
强大的	0.71		0.57	0.78		0.61			
理解的	0.70		0.58	0.72		0.55	0.70		0.51
成熟的	0.69		0.72				0.79		0.63
活跃的	0.69		0.64	0.75		0.62	0.70		0.53
镇静的	0.66		0.50	0.52		0.52	0.66		0.47
简单的(R)	−0.65		0.46				−0.67		0.45
方便的	0.65		0.49				0.78		0.64
神奇的	0.65		0.43	0.63		0.47			
开放的	0.64		0.59						
准确的	0.59		0.61						0.50
按部就班的(R)	−0.56		0.47				−0.71		0.65
值得信赖的	0.54		0.58	0.62		0.48			
聪明的							0.83		0.79
有趣的							0.68		0.53
宽厚的		0.88	0.79		0.90	0.82		0.91	0.88
低落的(R)		0.82	0.70		0.86	0.74		0.75	0.62
无恒心的(R)		0.81	0.67		0.86	0.74		0.74	0.57
孤僻的(R)		−0.80	0.66		−0.83	0.70		−0.75	0.60
缄默的(R)		−0.75	0.58		−0.76	0.58		−0.66	0.51
狡猾的(R)		0.73	0.57					0.64	0.47
动摇的(R)		0.70	0.53		0.74	0.57			
热情的		0.68	0.55		0.65	0.46		0.65	0.48
掩饰的(R)		0.65	0.44					0.64	0.44
奇怪的(R)		0.62	0.41					0.70	0.54
小心的		−0.56	0.44		−0.55	0.46			
特征值(旋转后)	12.83	6.27		10.13	5.12		11.21	4.82	
贡献率(旋转后)	35.62	17.41		37.51	18.96		38.67	16.63	
累计贡献率(旋转后)	35.62	53.04		37.51	56.47		38.67	55.30	

注：因子提取方法：最大似然法; 旋转法：具有 Kaiser 正态化最大方差法。R 表示反向计分。

通过探索性因子分析, 我们确定了两个主要因子, 第一个因子与能力相关, 第二个因子与热情相关。尽管男性和女性在词汇上的具体表现有所不同, 但两个群体均明显地显示出这两个主要因子。同时, 能力因子的方差解释力(35.62%)高于热情因子(17.41%)。

2.2.4 讨论

研究 1b 通过特质词评定任务对研究 1a 的结果进行了验证。根据统计分析结果, 人们对 LLMs 的感知主要集中于两个维度。第一个维度与智能、实用等能力词汇相关, 整体上反映了 LLMs 表现的能力特征。第二个维度则与宽厚、坦诚等友好词汇相关, 更多地反映了 LLMs 表现的热情特征。这一结果与研究 1a 的结果一致, 进一步验证了人们感知 LLMs 的主要维度是热情和能力。

值得注意的是, 部分词汇在因子归属上与“大二”模型存在一定差异。以“耐心”为例, 在人类特质评价中该特质通常归属于热情维度, 表征温暖关怀等情感属性(Abele et al., 2016); 然而在本研究中, “耐心”却被纳入能力维度。此现象可能表明, 当评价 LLMs 时, 人类倾向于将某些热情特质归类于算法性能而非情感属性。这种认知差异可能源于人类将其表现出的交互特征解读为程序设计的结果。此外, 源自大五人格严谨性维度的“小心的”词汇, 在本研究中为消极效价。这一效价反转可能由两方面原因导致。首先, 由于“大胆的”这一对照词缺失, 导致被试将“小心的”曲解为“畏忌的、有顾虑的”。其次, 由于 LLMs 需要通过精准响应规避风险, 因此“小心的”可能被解读为机械性回应而非审慎美德。同时, 值得注意的是, 条目“易用的”在能力因子上为负向载荷, 表明其可能反映了能力构念的反向特征, 即个体倾向于将高能力与较低的易用性联系在一起。总之, 因子分析显示, 归属于 LLMs 的热情和能力等特质词, 与归属于人类的特质词并不完全一致。

综上所述, 尽管人们对 LLMs 的感知主要围绕在热情和能力两个维度, 与对人类的感知类似, 但在细节上仍存在一些差异。因此, 未来应进一步系统构建适用于 LLMs 的热情与能力维度的特质词分类, 以从细节上检验热情和能力在 LLMs 和人类中的差异。

3 研究 2: 预测人类对 LLMs 的态度时感知维度的优先效应

3.1 研究目的

本研究旨在探讨人类对 LLMs 的感知如何影响人类的态度反应, 并进一步检验在预测不同态度时, 维度的优先效应是否会发生变化, 以验证假设 H3a 与 H3b。

3.2 方法

3.2.1 被试

为了确保被试能够准确表达个人对 LLMs 相关信息的评价, 我们在线招募具有 LLMs 使用经验的被试。与研究 1a 相同, 被试来自同一招募批次。在剔除 37 名不具备 LLMs 使用经验或未按指令作答的被试后, 最终有效被试人数为 178 名, 其中男性 105 名(58.9%), 女性 73 名(41.1%)。所有被试均自愿参与研究。

为了确保研究的样本量能够达到适当的统计检验力, 我们使用了 G*Power 软件基于回归分析(包含两个自变量和一个因变量)进行事前功效分析, 设定统计功效为 0.80, 效应量 f^2为 0.15 (代表中等效应), 显著性水平 α 为 0.05, 计算得出至少需要 68 名被试。本研究的研究样本量高于该预估样本量。

3.2.2 操作性定义

本研究通过主观权重检验优先效应, 即个体在评估时倾向赋予某一维度更高的重要性(Abele et al., 2021)。具体而言, 优先效应在本研究中被定义为: 预测人类对 LLMs 的态度时, 预测效力更高的维度被视为具有优先效应的维度。

3.2.3 工具与程序

为收集被试对 LLMs 的持续使用意愿和喜爱度等信息, 本研究采用问卷调查法。具体而言, 针对 LLMs 持续使用意愿的调查参考武晓宇和张大伟(2023)研究中使用的量表, 共 3 个条目, 分别为: (1)我愿意使用该 LLM; (2)我以后仍会使用该 LLM; (3)我愿意向他人推荐使用该 LLM。针对 LLMs 喜爱度的调查参考 Bartneck 等人(2009)编制的 Godspeed 量表中的喜爱度分量表, 共 5 个条目, 分别为: (1)这个 LLM 是讨人喜欢的; (2)这个 LLM 是亲切的; (3)这个 LLM 是待人礼貌的; (4)这个 LLM 是令人愉快的; (5)这个 LLM 是让人放心接近的。以上问卷均采用 7 点计分。为收集被试对 LLMs 的热情与能力的评分, 本研究采用佐斌等人(2018)根据 Abele 等人(2014)开发的测量刻板印象内容的特质词 7 点量表测量被试对人类和 LLMs 在热情和能力方面的评价。该量表包含热情和能力两个维度。其中, 热情特质词包括“热情的”、“友好的”; 能力特质词包括“聪明的”、“有能力的”。

被试首先选择自己较为熟悉的一个 LLM 进行

评价, 随后填写上述量表。

3.3 结果

3.3.1 共同方法偏差检验

我们采用 Harman 单因子检验共同方法偏差。具体来说, 我们将所有测量指标加载到一个单一潜变量上, 进行验证性因素分析(Confirmatory Factor Analysis, CFA)。如果存在共同方法偏差, 所有测量指标应会高度关联, 并且单一因子可以解释大部分的方差。然而, 结果显示, 单因子模型的拟合度较差($\chi^2(35) = 479.48$, $p < 0.001$, CFI = 0.69, TLI = 0.60, RMSEA = 0.28, SRMR = 0.17), CFI 和 TLI 这两项拟合指标低于 0.90, RMSEA 值高于 0.06, SRMR 高于 0.08, 表明模型与数据之间的拟合度较差。因此, 这一结果表明在本研究中不存在共同方法偏差。

3.3.2 描述统计

变量的相关性分析结果如表 4 所示。结果显示, 持续使用意愿、喜爱度、热情和能力之间均存在正相关关系。

3.3.3 回归分析

我们分别构建了两个线性回归模型, 分别以持续使用意愿和喜爱度为因变量, 热情和能力为自变量。结果显示, 热情和能力评分对所有因变量的预测效应均显著。具体而言, 持续使用意愿和喜爱度的 R^2分别为 28.59%和 30.64%, 调整后的 R^2分别为 27.77%和 29.85%, 具体结果见表 5 以及图 3、图 4。进一步分析发现, 在以持续使用意愿为因变量的模型中, 能力的解释效力高于热情; 而在以喜爱度为因变量的模型中, 热情的解释效力高于能力。此外, 我们构建了包含热情与能力交互项的模型, 但交互效应均未达到显著水平($p > 0.05$), 表明在本研究中热情与能力不存在交互效应。

表 4 各变量的描述性统计与相关系数矩阵(N = 178)

变量	*M*	*SD*	热情	能力	持续使用意愿	喜爱度
热情	5.21	1.24				
能力	5.24	1.18	0.30***			
持续使用意愿	5.96	1.17	0.33***	0.50***		
喜爱度	5.60	1.01	0.49***	0.39***	0.66***	

注: ***p<0.001

3.4 讨论

本研究支持假设 H3a 和 H3b, 结果表明, 被试对 LLMs 的热情和能力评价均能正向预测其持续使用意愿和喜爱度。在预测被试对 LLMs 的持续使用意愿时, 能力维度(β = 0.45, $p < 0.001$)相较于热情

表 5 被试对 LLMs 的热情和能力评分与对 LLMs 的持续使用意愿和喜爱度的多元回归分析结果

因变量	自变量	*b*	*SE*	β	*t*	*p*	95% CI
持续使用意愿	(截距)	2.70	0.41	–	6.61	< 0.001	[1.90, 3.51]
	热情	0.18	0.06	0.19	2.87	0.005	[0.06, 0.31]
	能力	0.44	0.07	0.45	6.64	< 0.001	[0.31, 0.57]
喜爱度	(截距)	2.65	0.35	–	7.62	< 0.001	[1.96, 3.33]
	热情	0.34	0.05	0.41	6.23	< 0.001	[0.23, 0.44]
	能力	0.23	0.06	0.27	4.05	< 0.001	[0.12, 0.34]

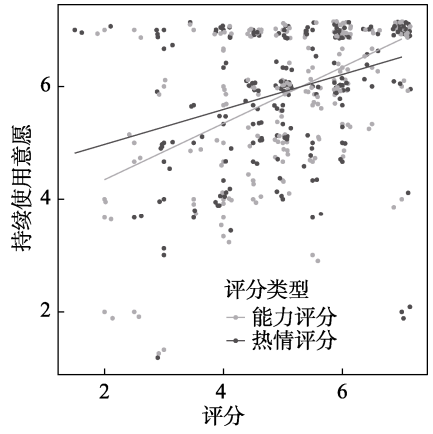

图 3 被试对 LLMs 的能力评分和热情评分与对 LLMs 持续使用意愿散点图

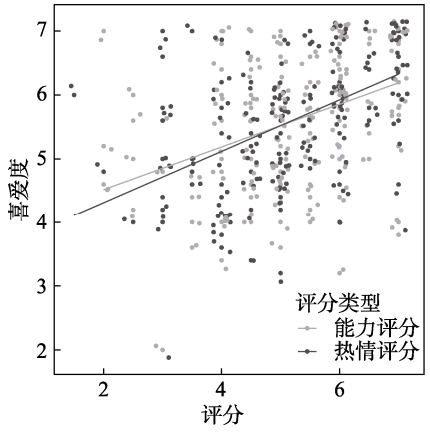

图 4 被试对 LLMs 的能力评分和热情评分与对 LLMs 喜爱度散点图

维度(β = 0.19, p = 0.005)具有更强的解释力。在预测被试对 LLMs 的喜爱度时, 热情维度(β = 0.41, p < 0.001)相较于能力维度(β = 0.27, p < 0.001)具有更强的解释力。

本研究揭示了人类对 LLMs 热情和能力的感知在预测人类对 LLMs 的不同态度时, 呈现出不同的优先效应。在预测功能型态度时, 能力维度更为重要; 而在预测情感型态度时, 热情维度更为重要。类似现象也出现在对人类的感知中。例如, 在职业环境中, 人们更倾向于依赖自身能力特征来描述自我, 而在家庭或社交情境中, 则更多依赖热情特征来描述自我(Uchronski, 2008)。本研究的发现提示, 态度类型可能是影响优先效应转换的关键因素。然而, 仍需注意的是, 态度通常是功能与情感混合的复合结构, 因此将持续使用意愿和喜爱度直接对应为功能型与情感型态度, 仍可能存在一定的简化。由于相较于态度类型, 情境类型的可控性较好, 同时, LLMs 已经被广泛应用于功能型场景(如提升工作效率、辅助任务完成)和关系型场景(如提供情感支持、陪伴互动)之中。因此, 未来研究有必要进一步研究情境类型对优先效应的影响, 深入理解优先效应在不同情境下的表现机制。

最后, 本研究结果对 LLMs 产品设计具有重要启示。对于功能型界面(如专业问答系统), 应突出准确性、响应速度等能力指标, 以提升人类对系统效能的信任和持续使用意愿。而对于情感陪伴型应用(如心理咨询机器人), 则应通过语言风格、交互节奏等方面强化热情感知, 以提升人类的情感依赖和喜爱度。

4　研究 3：人类对 LLMs 和对他人的热情和能力评分的比较研究

4.1　研究目的

本研究旨在比较人类对 LLMs 和对他人在热情与能力维度上的评价, 以检验 H4a 和 H4b。

4.2　方法

4.2.1　被试

本研究被试与研究 1a 为同一招募批次, 共招募 215 名被试, 排除其中 8 名完全没有 LLMs 使用经验和未按指令进行反应的被试, 最终有效被试人数为 207 名, 其中男性 124 名(59.90%), 女性 83 名(40.10%)。所有被试均自愿参与研究。为了确保研究的样本量能够达到适当的统计检验力, 我们使用了 G*Power 软件基于配对样本 t 检验进行统计检验力分析, 设定显著性水平 α 为 0.05, 效应量 d 为 0.50 (代表中等效应), 为了确保具有 80%的统计检验力, 至少需要 34 名被试。我们的研究样本量(N = 207)大于该预估样本量, 符合预定的统计检验力要求。

4.2.2　工具与程序

热情和能力感知量表与研究 2 一致。本研究要求被试对所接触过的大多数人以及所使用过的 LLMs 的热情和能力水平进行评价。

4.3　结果

配对样本 t 检验结果表明, 被试对 LLMs (M = 5.11, SD = 1.23)和对他人(M = 5.06, SD = 1.09)的热情评分没有显著差异, t(206) = 0.60, p = 0.551, Cohen's d = 0.05。但是, 被试对 LLMs (M = 5.16, SD = 1.20)和对他人(M = 4.81, SD = 1.23)的能力评分的差异显著, t(206) = 3.51, p < 0.001, Cohen's d = 0.29, 被试对 LLMs 的能力评价更高。具体结果见图 5 及表 6。

考虑到本研究中性别比例不平衡, 且男性和女性对人工智能的态度存在较大差异(Nomura et al., 2006), 我们进一步控制进行分析。结果显示, 男性被试对 LLMs (vs.人类)有更高的热情评价, t(123) = 2.30, p = 0.023, Cohen's d = 0.21; 对 LLMs (vs.人类)有更高的能力评价, t(123) = 2.28, p = 0.024, Cohen's d = 0.25。女性被试对 LLMs 和人类的热情评分差异不显著, t(82) = −1.15, p = 0.255, Cohen's d = −0.15; 对 LLMs (vs.人类)有更高的能力评价, t(82) = 2.71, p = 0.008, Cohen's d = 0.34。具体结果见表 6。

4.4　讨论

本研究结果表明, 被试对 LLMs 和对他人在热情维度上的评分没有显著差异, 因此未能支持假设 H4a; 然而, 在能力维度上, 被试对 LLMs 的能力评

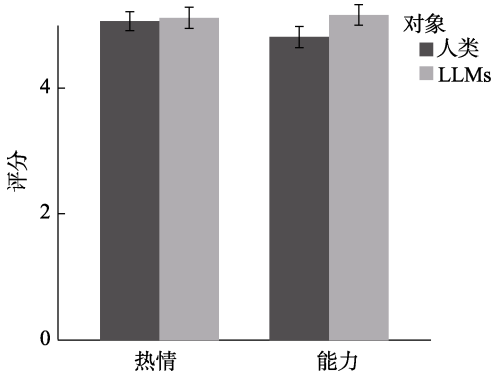

图 5　被试对人类和 LLMs 的热情和能力的评分

表 6 被试对 LLMs (vs.对人类)的热情和能力评分的配对样本 *t* 检验结果

组别	维度	*t*	*p*	*df*	95% CI	Cohen's *d*
全部(N = 207)	热情	0.60	0.551	206.00	[−0.12, 0.23]	0.05
	能力	3.51	<0.001	206.00	[0.15, 0.54]	0.29
男性(N = 124)	热情	2.30	0.023	123.00	[0.03, 0.40]	0.21
	能力	2.28	0.024	123.00	[0.04, 0.53]	0.25
女性(N = 83)	热情	−1.15	0.255	82.00	[−0.52, 0.14]	−0.15
	能力	2.71	0.008	82.00	[0.12, 0.77]	0.34

分显著高于对人类的评分, 支持假设 H4b。这一发现揭示了在当前技术水平下, 在人类的感知中, LLMs 在能力表现上相较于人类被赋予更高的评价; 同时, 尽管其在热情维度上的表现未能超越人类, 但总体与人类一般水平一致。因此, 从整体印象来看, 如果将其与 Durante 等人(2017)所研究的群体刻板印象相比照, 我们发现在本研究中, LLMs 在人类感知中所呈现的高能力、中等热情的特征, 与高度和平国家(如丹麦)中, 人们对高学历群体、精英、富人以及男性的印象相似。从既有的研究来看, LLMs 在热情和能力两个维度的确具有良好表现。尽管尚未有直接测量 LLMs 热情的相关研究, 但有研究表明 LLMs 的宜人性和共情能力表现较好(Bodroža et al., 2024; Sorin et al., 2024), 甚至有研究表明 LLMs 在共情方面的表现超过了人类(Lee et al., 2024)。宜人性通常与信任他人、坦诚交往、助人、谦逊、富有同情心等相关(Heaven et al., 2013), 而共情能力指的是体验他人内在情感的能力(Cuff et al., 2016)。因此, 宜人性和共情的内涵与热情有一定的重合, 均体现了友好、关怀等特点。其次, 在能力维度上, 当前对 LLMs 的标准化能力测评结果提供了有力证据, 如研究发现, LLMs 在多项能力测试中的表现突出(Bubeck et al., 2023; Webb et al., 2023)。

此外, 性别因素在本研究中的影响也值得关注。本研究发现, 男性被试对 LLMs 的热情评分显著高于人类, 而女性群体未表现出显著差异。这一结果与先前关于性别与对人工智能感知的关系的研究一致(Nomura et al., 2006), 表明性别差异可能在人类对人工智能的评价中起到重要作用。这可能是由于女性在技术使用中更依赖情感反馈, 而男性则可能更多依赖任务的完成度和技术的实际效用(Venkatesh & Morris, 2000)。因而在感知 LLMs 的过程中, 女性可能会因为对技术的情感需求(例如语言生成的互动内容、情感回应)而表现出更多的情感投入, 更敏锐地感知人工智能在热情维度的表现; 而男性则可能更多关注技术的功能性, 对热情的感知满足于“够用即好”。

5 总讨论

本研究基于新一代人工智能体(如 LLMs)在人类社会深度融合的背景下, 旨在从人类感知的视角探讨人机互动的特征。通过构建人类感知 LLMs 的理论框架, 为人机共生时代下人类与人工智能关系的理论建构与实证研究提供支持。

本研究以社会认知内容的“大二”模型为基础, 通过三项研究探讨热情与能力维度在人类感知 LLMs 中的作用。研究 1 的结果表明, 热情和能力是人类感知 LLMs 的主要维度, 且在一般情境下, 能力优先于热情。研究 2 发现, 人类对 LLMs 热情和能力的评价均能正向预测人类对 LLMs 的持续使用意愿和喜爱度, 其中能力维度对人类对 LLMs 的持续使用意愿的预测效力较大, 而热情维度则对喜爱度的预测效力较大。研究 3 的结果表明, 在人类的感知中, LLMs 在能力维度超越了人类一般水平, 而热情维度与人类一般水平不存在显著差异。总体而言, 人类感知 LLMs 的框架与感知他人相似, 均主要围绕在热情和能力两个维度, 但是, 在一般情境下, 能力优先于热情。此外, LLMs 在人类的感知中更接近于高能力、中等热情的个体, 而非人类的一般水平。

5.1 理论贡献

本研究的核心理论贡献体现在两个方面：首先, 扩展了“大二”模型的适用范围; 其次, 拓宽了人机交互领域的理论视角。

首先, 本研究初步构建了人类感知 LLMs 的理论框架, 拓展了“大二”模型的适用范围, 将感知对象从人类拓展至 LLMs。一方面, 就感知维度而言, Fiske 等人(2007)从进化论的角度解释热情和能力两个维度作为刻板印象内容的主要维度的合理性, 指出二者分别对应意图检测和攻击能力检测两个方面, 因此对二者的关注有助于保障生存(Fiske

et al., 2007)。类似地, 双视角模型通过功能性的角度来解释两个维度的合理性, 提出社会认知是服务于感知者的目标(Abele &Wojciszke, 2014), 而这种目标相较于进化论中的生存目标会更加具体。无论何种解释, 在 LLMs 技术快速发展的背景下, 它可能给人们造成潜在威胁并影响着人们目标的达成, 因此热情与能力仍然是感知 LLMs 的主要维度。另一方面, 本研究对维度的优先效应进行了初步探索。研究表明, 在一般情境下, 与对人类感知中的热情优先不同, 对 LLMs 感知中能力优先。正如双视角模型所提出优先效应源于利益最大化原则, 在感知 LLMs 时仍然遵循该原则。利益最大化原则是指, 在感知他人时, 由于热情作为一种他人有利特质, 为实现自我利益的最大化, 人们会优先关注(Abele &Wojciszke, 2007)。而在感知 LLMs 时, 热情和能力均是他人有利特质, 且能力的利他性会更直接, 因此在感知 LLMs 时能力优先。然而, 在对态度预测中, 优先效应会发生变化。具体而言, 在预测功能型态度(即持续使用意愿)时, 能力优先; 而在预测情感型态度(即喜爱度)时, 热情优先。类似的优先效应转换现象在对人类的感知中同样会出现, 例如研究发现在不同的情境下以及对不同的感知对象, 优先效应会发生变化(Abele &Wojciszke, 2007; Uchronski, 2008; Wojciszke& Abele, 2008)。

其次, 本研究拓宽了人机交互领域的理论, 尤其是拓展了用户体验相关领域的理论基础。以往关于人工智能用户体验的研究, 主要依赖技术接受模型(Technology Acceptance Model, TAM)和技术接受与使用统一理论(Unified Theory of Acceptance and Use of Technology, UTAUT), 重点关注绩效预期、易用性等传统因素对用户使用态度的影响(Davis, 1989; de Ruyter et al., 2005; Venkatesh et al., 2003)。近年来, 随着技术的发展, 研究者逐步拓展了影响人类对人工智能态度的变量, 包含来自情境、用户特征、机器特征等多方面的因素。例如, 研究发现, 人们对机器人的信任倾向、机器人具备的可信功能与设计, 以及所处的服务任务和情境, 会影响用户对社会服务机器人的互动信任(Chi et al., 2021)。然而, 现有研究在不断扩展变量的过程中, 使相关理论模型愈加复杂, 尚未形成一个更加简洁而系统的解释框架。本研究在此基础上进一步拓展了人机交互的理论视角, 将热情与能力两个核心社会认知内容维度引入人类与 LLMs 交互的研究中, 系统探讨了其对用户态度的影响。这一理论为人机交互研究提供了新的视角, 有助于优化人机互动方式, 提升系统的可控性与安全性。

5.2　实践意义

基于上述两方面主要理论贡献, 本研究在社会和技术两方面也具备相应的实践意义。

首先, 在社会实践方面, 本研究将 LLMs 纳入“大二”模型框架后, 为人类社会设计人机共生时代更合理的制度安排和伦理规范提供了重要的理论依据和参考方向。在制度安排方面, 基于 LLMs 在热情和能力方面的优良表现, 政策制定者需要考虑如何将 LLMs 纳入现有的法律体系, 如制定专门的 LLMs 法案, 明确 LLMs 的法律地位、权利与义务, 以及监管机制, 确保 LLMs 的发展与应用不违背社会公共利益。同时, 随着 LLMs 能力的提升, 其对就业市场的影响日益显著。本研究鼓励从“人与 LLMs 共生”的角度出发, 设计促进人机协作的就业政策, 如提供技能培训、转型援助等, 以减少技术进步带来的社会不平等。在伦理与社会责任相关设计方面, 本研究提醒开发者在追求技术进步的同时, 需关注 LLMs 的伦理和社会责任。通过合理设计 LLMs 的行为模式和交互方式, 确保其不会对社会造成负面影响, 而是成为促进人类福祉的有力工具。

其次, 在技术实践方面, 本研究在人工智能领域, 特别是针对 LLMs 的技术实践具有重要意义, 为 LLMs 的设计、开发与应用提供了新的理论基础。在增强 LLMs 的可解释性与可控性方面, 本研究通过构建人类感知 LLMs 的理论模型, 揭示了热情与能力维度在 LLMs 感知中的重要性, 为理解和控制 LLMs 的行为提供了新框架。这不仅有助于开发者设计更易于解释和调控的算法, 减少 LLMs 行为的不可预测性, 从而提高其在社会应用中的安全性和可靠性。还有助于 LLMs 进行技术优化, 开发者可以聚焦于提升 LLMs 在这两个维度上的表现, 如通过增强对话的自然流畅性来提升能力感知, 或通过设计更具同理心的回复来增强热情感知, 从而使用户更容易接受和信任 LLMs。同时, 开发者可以根据不同的应用场景和需求, 设计具有特定社会角色的 LLMs, 如教育助手、心理陪伴者等, 以更好地服务于人类社会。

5.3　研究局限与未来展望

第一, 从感知维度到切面的拓展。近年来, 社会认知内容的研究越来越关注热情与能力的切面, 以解释不同理论模型之间的差异, 并进一步探索人

类心理和行为的机制(Abele et al., 2021)。具体而言, 双视角模型将热情分为道德与友好两个切面, 能力分为才智与自信两个切面(Abele et al., 2016)。在本研究中, 研究 1a 发现, 对 LLMs 热情感知主要集中在道德层面, 对能力感知主要集中在才智层面; 然而, 在研究 1b 中未表现出到明显的切面差异。这一差异可能与研究 1b 中评价词汇较少有关。此外, 如在研究 1b 中“耐心”被视为 LLMs 能力的一部分, 亦提示其能力维度可能涵盖了不同于人类的特殊外延。未来可通过爬虫等技术, 从自然语境中提取更丰富的、真实的描述性词汇, 扩大词汇库, 探索热情与能力切面在感知 LLMs 中的适用性, 完善理论框架, 细化两维度内涵与边界。

第二, 维度的优先效应的场景化、动态化分析。我们发现维度的优先效应在态度预测中会发生转换, 随着 LLMs 从提升效率到提供情感支持等多种情境下的广泛应用, 我们需要深入研究维度优先效应在不同情境中的表现。未来研究可以通过实验的方式进一步探讨其产生机制和转化条件, 可能的关键因素包括情境类型(如功能型情境与情感型情境)。

第三, 探索影响人类对 LLMs 感知的复杂机制。本研究没有对影响人类对 LLMs 感知的线索(即原因)和影响机制进行深入研究。未来研究可以进一步细化, 构建完整的人类对 LLMs 的感知理论。

最后, 本研究促使我们在社会层面重新审视“人类”范畴。本研究旨在构建人类对 LLMs 感知的理论框架, 为人类与 AI 共生时代的社会研究提供基础。本研究仅局限于探讨在人类的感知中 LLMs 的形象, 并将其与人类对他人的感知相比较, 未涉及 LLMs 本体的类人性这样更加深刻的命题。但是, 无论是 LLMs 在人类感知中的热情与能力表现与人类的相似, 还是 LLMs 本体的类人性的演化, 都不可避免地改变 LLMs 在人机共生社会中所扮演的社会角色, 从而可能引发公众对人机边界的模糊认知。具体而言, 当前 LLMs 功能日益完善, 开发者面临两难：一方面赋予其类人特质, 另一方面又以伦理为由施加限制, 强调其工具属性。然而, 人类在特定场景下的人机交互中仍可能倾向于将 LLMs 视为近似“人”的存在。更深层的问题是, 随着技术进步, “人类”概念的边界是否在社会学意义上发生变化？这一问题不仅关涉对人工智能的社会角色认知, 也触及人类自我认知。若 LLMs 被视为具备人类心智能力, 在社会层面, “人类”定义是否已被泛化?例如, 近期研究表明, 借助人类感知层面的同一性, 可以将本体对等性拓展至非人类元素, 这也提示了“人类”概念在社会认知中的可塑性(Sundararajan et al., 2022)。这一争议尚存, 未来研究可借助心灵知觉理论(Mind Perception Theory), 从经验性和能动性维度深入研究 LLMs 的本体类人性(Gray et al., 2007; Gray & Wegner, 2010)。该理论虽与“大二”模型部分重叠(Heflick et al., 2011; Waytz et al., 2010; Waytz& Young, 2014), 但其更强调“心灵”的认知构建。其中, 经验性指代个体的情感和感知能力, 能动性则涉及自主行动与自我控制能力, 若缺失任一维度, 则该实体在认知层面上难以被赋予“心灵”——这一传统上被视为人类独有的特征。因此, 对 LLMs 本体类人性的探讨及其在社会互动中的表现, 或将在未来见证“人类”范畴的深刻演变, 并进一步推动关于人机共生社会的理论构建。

6 结论

本研究主要结论如下：第一, 热情和能力是人类感知 LLMs 的主要维度, 且在一般情境中, 能力较之热情具有优先效应; 第二, 人类对 LLMs 的热情和能力评分均能正向预测人类对 LLMs 的持续使用意愿和喜爱度, 但其中能力对持续使用意愿的预测效力较高, 而热情对喜爱度的预测效力较强; 第三, 在人类的感知中, 对 LLMs 的热情评分与对他人的热情评分没有显著差异, 但对 LLMs 的能力评分显著高于对一般人类的能力评分。

参 考 文 献

Abele, A. E., & Brack, S. (2013). Preference for other persons' traits is dependent on the kind of social relationship. *Social Psychology*, *44*(2), 84−94.

Abele, A. E., Bruckmüller, S., & Wojciszke, B. (2014). You are so kind − and I am kind and smart: Actor − observer differences in the interpretation of on-going behavior. *Polish Psychological Bulletin; 45*(4), 394−401.

Abele, A. E., Ellemers, N., Fiske, S. T., Koch, A., & Yzerbyt, V. (2021). Navigating the social world: Toward an integrated framework for evaluating self, individuals, and groups. *Psychological Review*, *128*(2), 290−314.

Abele, A. E., Hauke, N., Peters, K., Louvet, E., Szymkow, A., & Duan, Y. (2016). Facets of the fundamental content dimensions: Agency with competence and assertiveness-communion with warmth and morality. *Frontiers in Psychology*, *7*(1), Article 1810.

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology*, *93*(5), 751−763.

Abele, A. E., &Wojciszke, B. (2014). Communal and agentic

content in social cognition: A dual perspective model. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol 50, pp. 195−255). Elsevier Academic Press Inc.

Abele, A. E., &Wojciszke, B. (Eds). (2018). *Agency and communion in social psychology*. Routledge.

Bai, X., Ramos, M. R., & Fiske, S. T. (2020). As diversity increases, people paradoxically perceive social groups as more similar. *Proceedings of the National Academy of Sciences*, *117*(23), 12741−12749.

Bartneck, C., Kulic, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, *1*(1), 71−81.

Bodroža, B., Dinić, B. M., & Bojić, L. (2024). Personality testing of large language models: Limited temporal stability, but highlighted prosociality. *Royal Society Open Science*, *11*(10), 240180.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (pp. 1877−1901). Curran Associates Inc.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv*. https://doi.org/10.48550/arXiv.2303.12712

Carpinella, C. M., Wyman, A. B., Perez, M. A., & Stroessner, S. J. (2017, March). The Robotic Social Attributes Scale (RoSAS): Development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 254−262). ACM.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245−276.

Chi, O. H., Jia, S., Li, Y., & Gursoy, D. (2021). Developing a formative scale to measure consumers' trust toward interaction with artificially intelligent (AI) social robots in service delivery. *Computers in Human Behavior*, *118*, 106700.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed. pp. xii, 430). Lawrence Erlbaum Associates, Inc.

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, *10*(1), Article 7.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology*, *92*(4), 631−648.

Cuddy, A. J. C., Fiske, S. T., Kwan, V. S. Y., Glick, P., Demoulin, S., Leyens, J. P., ... Ziegler, R. (2009). Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology*, *48*(1), 1−33.

Cuff, B. M. P., Brown, S. J., Taylor, L., & Howat, D. J. (2016). Empathy: A review of the concept. *Emotion Review*, *8*(2), 144−153.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319−340.

de Ruyter, B., Saini, P., Markopoulos, P., & van Breemen, A. (2005). Assessing the effects of building social intelligence in a robotic interface for the home. *Interacting with Computers*, *17*(5), 522−541.

dos Santos, R. P. (2023). Enhancing physics learning with ChatGPT, Bing Chat, and Bard as agents-to-think-with: A comparative case study.*arXiv*. https://doi.org/10.48550/arXiv.2306.00724

Durante, F., Fiske, S. T., Gelfand, M. J., Crippa, F., Suttora, C., Stillwell, A., ... Teymoori, A. (2017). Ambivalent stereotypes link to peace, conflict, and inequality across 38 nations. *Proceedings of the National Academy of Sciences*, *114*(4), 669−674.

Fiske, S. T. (1992). Thinking is for doing: Portraits of social cognition from daguerreotype to laserphoto. *Journal of Personality and Social Psychology*, *63*(6), 877−889.

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, *27*(2), 67−73.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77−83.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878−902.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619.

Gray, K., & Wegner, D. M. (2010). Blaming God for our pain: Human suffering and the divine mind. *Personality and Social Psychology Review*, *14*(1), 7−16.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493−498.

Heaven, P. C. L., Ciarrochi, J., Leeson, P., & Barkus, E. (2013). Agreeableness, conscientiousness, and psychoticism: Distinctive influences of three personality dimensions in adolescence. *British Journal of Psychology*, *104*(4), 481−494.

Heflick, N. A., Goldenberg, J. L., Cooper, D. P., & Puvia, E. (2011). From women to objects: Appearance focus, target gender, and perceptions of warmth, morality and competence. *Journal of Experimental Social Psychology*, *47*(3), 572−581.

Huang, J., Wang, W., Lam, M. H., Li, E. J., Jiao, W., & Lyu, M. R. (2023). Revisiting the reliability of psychological scales on large language models. *arXiv*. https://doi.org/10.48550/arXiv.2305.19926

Jiang, G. Y., Xu, M. J., Zhu, S. C., Han, W. J., Zhang, C., & Zhu, Y. X. (2023, December). Evaluating and inducing personality in pre-trained language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Neural Information Processing Systems (NIPS).

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, *121*(45), e2405460121.

Lee, Y. K., Suh, J., Zhan, H., Li, J. J., & Ong, D. C. (2024). Large language models produce responses perceived to be empathic. *arXiv*. https://doi.org/10.48550/arXiv.2403.18148

Luo, J., & Dai, X. Y. (2018). Development of the Chinese Adjectives Scale of Big-Five Factor Personality Ⅳ: A Short Scale Version. *Chinese Journal of Clinical Psychology*, *26*(4), 642−646.

[罗杰, 戴晓阳. (2018). 中文形容词大五人格量表的初步编制Ⅳ: 简式版的研制. *中国临床心理学杂志*, *26*(4),

642−646.]
MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84−99.
McKee, K. R., Bai, X. C. Z., & Fiske, S. T. (2023). Humans perceive warmth and competence in artificial intelligence. *Iscience*, *26*(8), 107256.
Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., & Tscheligi, M. (2017). To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI*, *4*(1), Article 21.
Mohammad, Y., & Nishida, T. (2015). Why should we imitate robots? Effect of back imitation on judgment of imitative skill. *International Journal of Social Robotics*, *7*(4), 497−512.
Mori, M. (1970). The uncanny valley. *Energy*, *7*, 33−35.
Nicolas, G., Bai, X. C. Z., & Fiske, S. T. (2021). Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*, *51*(1), 178−196.
Nomura, T., Kanda, T., & Suzuki, T. (2006). Experimental investigation into influence of negative attitudes toward robots on human−robot interaction. *Ai & Society*, *20*(2), 138−150.
Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*(27730−27744), Article 2011.
Scheunemann, M. M., Cuijpers, R. H., & Salge, C. (2020). Warmth and competence to predict human preference of robot behavior in physical human-robot interaction. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (pp. 1340−1347). 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN).
Sorin, V., Brin, D., Barash, Y., Konen, E., Charney, A., Nadkarni, G., & Klang, E. (2024). Large language models and empathy: Systematic review. *Journal of Medical Internet Research*, *26*(10), Article e52597.
Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., ... Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, *8*(7), 1285−1295.
Sundararajan, L., Wu, M. S., Ho, W.-T., Sun, J. C., Leung, C. P., & Rich, G. J. (2022). Expanding our kind: A pan-cultural study of the animistic principle of ontological parity. *The Humanistic Psychologist*, *50*(3), 476−496.
Tae, M., & Lee, J. (2020). The effect of robot's ice-breaking humor on likeability and future contact intentions. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 462−464). ACM.
Uchronski, M. (2008). Agency and communion in spontaneous self-descriptions: Occurrence and situational malleability. *European Journal of Social Psychology*, *38*(7), 1093−1102.
Venkatesh, V., & Morris, M. (2000). Why don't men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behavior. *MIS Quarterly*, *24*(1), 115−139.
Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425−478.
Wang, J., Ma, W., Sun, P., Zhang, M., & Nie, J.-Y. (2024). Understanding user experience in large language model interactions. *arXiv*. https://doi.org/10.48550/arXiv.2401.08329
Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219−232.
Waytz, A., & Young, L. (2014). Two motivations for two dimensions of mind. *Journal of Experimental Social Psychology*, *55*, 278−283.
Webb, T., Holyoak, K. J., & Lu, H. J. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526−1541.
Wojciszke, B., & Abele, A. E. (2008). The primacy of communion over agency and its reversals in evaluations. *European Journal of Social Psychology*, *38*(7), 1139−1147.
Wu, X. Y., & Zhang, D. W. (2023). The impact mechanism of mobile visual search continuous behavior intention from academic new media users: Empirical analysis of UTAUT2-ECT integration model. *Journal of Intelligence*, *42*(8), 164−176.
[武晓宇, 张大伟. (2023). 学术用户移动视觉搜索的持续使用意愿研究——基于UTAUT2-ECT整合模型的实证分析. *情报杂志*, *42*(8), 164−176.]
Zhao, Y. K., Huang, Z., Seligman, M., & Peng, K. P. (2024). Risk and prosocial behavioural cues elicit human-like response patterns from AI chatbots. *Scientific Reports*, *14*(1), Article 7095.
Zuo, B., Dai, T. T., Wen, F. F., &Suo, Y. X. (2015). The big two model in social cognition. *Journal of Psychological Science*, *38*(4), 1019−1023.
[佐斌, 代涛涛, 温芳芳, 索玉贤. (2015). 社会认知内容的"大二"模型. *心理科学*, *38*(4), 1019−1023.]
Zuo, B., Wen, F. F., Wu, Y., &Dai, T. T. (2018). Situational evolution of the relationship between warmth and competence in intergroup evaluation: Impact of evaluating intention and behavioral outcomes. *ActaPsychologicaSinica*, *50*(10), 1180−1196.
[佐斌, 温芳芳, 吴漾, 代涛涛. (2018). 群际评价中热情与能力关系的情境演变:评价意图与结果的作用. *心理学报*, *50*(10), 1180−1196.]

Humans perceive warmth and competence in large language models

WU Yueting[1], WANG Bo[2,3], BAO Han Wu Shuang[4], LI Ruonan[1],
WU Yi[1], WANG Jiaqi[1], CHENG Cheng[3], YANG Li[3,5]

([1] *School of Education, Tianjin University, Tianjin* 300354, *China*)
([2] *College of Intelligence and Computing, Tianjin University, Tianjin* 300354, *China*)
([3] *Institute of Applied Psychology, Tianjin University, Tianjin* 300354, *China*)
([4] *School of Psychology and Cognitive Science, East China Normal University, Shanghai* 200062, *China*)
([5] *Laboratory of Suicidal Behavior Research, Tianjin University, Tianjin* 300354, *China*)

Abstract

The rapid development and application of Large Language Models (LLMs) have significantly enhanced their capabilities, influencing human-machine interactions in profound ways. As LLMs evolve, society is shifting from traditional interpersonal interactions to a multilayered structure integrating human-to-human, human-to-machine, and machine-to-machine interactions. In this context, understanding how humans perceive and evaluate LLMs—and whether this follows the Big Two model of warmth and competence in interpersonal perception—has become critical. This study examines human perceptions of LLMs through three progressive empirical studies.

Participants with prior LLM experience were recruited for the studies. Study 1 comprised two sub-studies: Study 1a (N = 207) used a free-response task, asking participants to describe their impressions of LLMs using at least three words, which were analyzed using the *Semi-Automated Dictionary Creation for Analyzing Text* to identify key dimensions of perception. Study 1b (N = 219) involved a lexical rating task, in which participants rated the applicability of selected evaluation words to LLMs. Study 2 (N = 178) used a questionnaire, in which participants rated a familiar LLM and provided feedback on their willingness to continue using it and their liking of it. Study 3 (N = 207) employed a questionnaire survey to assess participants' ratings of warmth and competence for both humans and LLMs.

Study 1 found that humans primarily perceive LLMs through warmth and competence, similar to how they perceive other humans. In general contexts, participants prioritized competence over warmth when evaluating LLMs, showing a significant priority effect (odds ratio = 2.88, z = 9.512, 95% CI [2.32, 3.59], p< 0.001). This contrasts with the typical warmth-priority effect in human-to-human perception. Study 2 investigated the relationship between perceptions of warmth and competence and human attitudes toward LLMs, specifically their emotional (e.g., liking) and functional (e.g., willingness to continue using) attitudes. Results showed that both dimensions positively predicted participants' liking and willingness to continue using LLMs. Warmth had a stronger predictive effect on liking (warmth: β = 0.41, p < 0.001; competence: β = 0.27, p < 0.001), while competence had a stronger predictive effect on willingness to continue using (warmth: β = 0.19, p = 0.005; competence: β = 0.45, p < 0.001). This outcome suggests that the priority effect of warmth and competence shifts across attitude predictions. Study 3 examined specific LLMs ratings in terms of warmth and competence. Results showed no significant difference in warmth ratings between humans (M = 5.06, SD = 1.09) and LLMs (M = 5.11, SD = 1.23), t(206) = −0.60, p = 0.551. However, LLMs were rated significantly higher on competence (M = 5.16, SD = 1.20) than humans (M = 4.81, SD = 1.23), t(206) = −3.51, p < 0.001, Cohen's d = −0.29.

This study makes two significant contributions to the field. First, it establishes a preliminary theoretical framework for understanding human perception of LLMs. Second, it offers new insights into human-machine interaction by emphasizing the importance of warmth and competence in shaping user attitudes. The findings have practical implications for AI design and policymaking, providing a framework for improving user acceptance, optimizing LLM design, and promoting responsible human-AI coexistence.

Keywords　large language model, social cognition, warmth, competence