心理学报 2025, Vol. 57, No. 11, 1951–1972 https://doi.org/10.3724/SP.J.1041.2025.1951
Acta Psychologica Sinica https://cstr.cn/32110.14.2025.1951

人与 AI 对智能家居机器人的安全信任及其影响因素*

由姗姗[1,2] 齐 玥[1,2] 陈俊廷[1,2] 骆 磊[1,2] 张 侃[3,4]

([1]中国人民大学心理学系; [2]中国人民大学心理学系实验室, 北京 100872)
([3]中国科学院院心理研究所认知科学与心理健康全国重点实验室, 北京 100101)
([4]中国科学院大学心理学系, 北京 100049)

摘 要 随着智能家居机器人技术的发展, 安全风险成为人机信任的新挑战。本研究提出并验证了智能家居机器人信任的新维度——安全信任。为此, 研究 1 编制了智能家居机器人安全信任量表, 并验证了人机信任三因子结构的稳定性和信效度。研究 2 和研究 3, 深入分析了机器人的静态和动态特征对人类与人工智能(AI)使用者安全信任的影响。结果发现, 在静态特征上, 人们对身高较矮以及摄像头不明显的机器人安全信任水平更高; 并且机器人拟人化程度影响了人类对这些静态特征的敏感性。在动态特征上, 机器人较慢的运动速度和摄像头关闭动作提高了人类的安全信任, 同时, 不同场景下这些动态特征的影响存在差异。此外, AI 与人类在安全信任上表现出一定的一致性, 但总体上 AI 对机器人摄像头的敏感度低于人类。本研究结果为家居机器人的设计与制造提供了重要的理论支持和实践指导。

关键词 人机信任, 安全信任, 智能家居机器人, 使用意愿, 大语言模型

分类号 B849: C91

1 前言

在人工智能(Artificial Intelligence, AI)时代, 机器人逐渐成为人类社会活动中的重要参与者, 在教育(Leyzberg et al., 2014)、医疗(Srinivasan et al., 2022)、商业(Nawaz, 2019)等多个领域发挥重要作用。在家用场景中, 智能家居机器人的应用也在逐渐普及(Cagiltay & Mutlu, 2024)。比如机器人 Kuavo 曾在 2024 中国家电及消费电子博览会(AWE)上展示了其在家居环境中的洗衣、园艺等功能(AWE, 2024)。除了完成清扫房间等家务, Jibo 等机器人还可以通过感知和交互技术, 为家庭成员提供情感上的陪伴与支持(Rane et al., 2014)。伴随技术的进步, 机器人的智能化水平不断提升, 人与机器人的关系从人与机器的工具关系, 逐渐转向人与智能的社会关系(许为, 葛列众, 2020; 许为 等, 2024)。人们对于机器人的信任也随着人机关系的演变而发生变化, 逐渐接近人际信任(Lewis & Marsh, 2022; Mayer et al., 1995)。在人机交互过程中, 使用者的信任是人们是否使用机器人的重要决定因素(You & Robert, 2018)。对于智能家居机器人信任的研究, 不仅有助于促进其应用与推广, 也是智能时代人机关系研究中的重要议题。

人工智能与传统技术之间的显著差异, 催生了许多新的关注点, 深化了人机交互的复杂性。例如, 智能系统的人格感知(如 AI 的个性化特征)、人与智能系统关系联结和团队协作等问题, 引发了对人机交互的新探讨(Li & Qi, 2025; Mei et al., 2024; Xu et al., 2024; Zhang et al., 2023)。为了适应日益复杂的人机交互环境, 信任研究需要进一步深化。在智

收稿日期: 2024-08-26

* 国家自然科学基金(32471130, 32000771)、中国人民大学科学研究基金(中央高校基本科研业务费专项资金资助)项目成果(21XNLG13)、2018 年度中央高校建设世界一流大学(学科)和特色发展引导专项资金(RUCPSY0007)支持。

通信作者: 齐玥, E-mail: qiy@ruc.edu.cn

能家居这一特定应用场景中, 智能机器人的自主活动与家庭成员的日常生活和隐私保护紧密相关, 这也使得机器人的高智能性可能引发用户新的担忧。以往的研究虽然已经证实了安全性感知在人机交互中的重要性, 但尚未深入探讨安全感知为人机信任带来的潜在危机(Bartneck, 2023; Zacharaki et al., 2020)。因此, 本研究提出并验证了智能家居机器人信任的新维度——安全信任。从理论层面来看, 安全信任的提出扩展和完善了现有的信任模型, 为新时代人机交互中的信任研究提供新的视角; 从实践层面来看, 深入研究安全信任能够为智能家居机器人的设计与优化提供有价值的指导, 进而促进其在家庭生活中的广泛应用。

与此同时, AI 的角色也逐渐从传统的任务执行者扩展到决策主体, 特别是在多智能体协作系统中, AI 不仅需要作为任务执行者承担责任, 还需要作为信任者参与决策过程(齐玥 等, 2024; Xie et al., 2024)。这一变化不仅要求 AI 能够理解和应对人类的信任期望, 还要求其在决策过程中展现出一致性和透明度。因此, 探索 AI 的信任模式, 尤其是在复杂家庭环境中的信任机制, 有助于我们更好地理解和设计人机合作的深层次互动方式, 并为未来智能家居机器人在更多场景中的应用提供理论支持。

1.1 智能家居机器人

智能家居机器人属于服务机器人的一种, 是指在家庭环境中为用户提供服务、拥有一定感知、交互、学习能力的自动化机器人(孙效华 等, 2020)。不同于最先出现的清洁机器人, 如今的家居机器人已逐步实现了多样化的功能(Prassler et al., 2016)。比如, Kao 和 Wang (2015)设计的家居机器人拥有对话交互、拍照、远程监控、定时提醒的功能。陪伴机器人 Buddy 可以巡视房屋, 发现隐患时及时报警, 同时也可以和其他智能家居产品整合起来, 获得远程操控家电的能力(Milliez, 2018)。现在, 智能家居机器人的功能多样, 并且随着技术进步不断发展, 因此有必要确定研究所讨论的范围。Lee (2021)对服务机器人领域进行了系统的文献综述, 将服务机器人进行了分类(专业非社交型、专业社交型、家庭/个人非社交型和家庭/个人社交型四种类型), 并总结了服务机器人的关键技术领域(抓取、检测、导航、人机交互和架构/平台等)。基于此, 本文所讨论的智能家居机器人属于家庭/个人社交型服务机器人, 使用场景在家庭中, 拥有抓取、检测、导航和信息存储等功能, 能够与用户进行交互, 并且以独立架构而非平台的形式呈现(Wan et al., 2016)。智能家居机器人即将进入越来越多的普通家庭, 研究使用者对其的信任将为智能家居机器人的推广和应用奠定理论基础。

1.2 人类对智能家居机器人的安全信任

人机信任是在已知不确定和脆弱的情况下, 认为代理(agent)能帮助个体实现目标的态度(Lee & See, 2004)。Billings 等人回顾了 302 个信任定义, 包括 220 个人际信任定义以及 82 个自动化信任定义, 发现大量的自动化信任定义涉及到用户对自动化系统的期望、信心、风险、脆弱性、依赖、态度及合作等特征(Billings et al., 2012)。这些信任定义揭示了人机信任的核心特征。虽然在定性研究中人们很少会报告是因为信任因素才选择使用机器人, 但人机信任已在多个研究中被证实可以影响人们对机器人的使用选择(Robinette et al., 2017; Sanders et al., 2019)。随着技术的不断进步, 人机信任的内涵从对基础工具使用的信任(例如机械设备), 到对具备一定自主性系统的信任(如导航系统), 再到如今对具备复杂交互能力的 AI 系统的信任, 其维度逐步丰富和深化(齐玥 等, 2024)。

人机信任与人际信任有一定相似性, 例如两者都涉及信任者对信任对象可靠性、动机和能力的评估(Lewis & Marsh, 2022; Mayer et al., 1995), 但二者在多个方面仍然表现出显著的差异。人际信任通常基于情感联结和社会互动经验, 而人机信任更多依赖于对技术功能的理性评估。此外, 人际信任是双向的, 而人机信任更多表现为单向信任, 即人类对机器的信任。然而, 人机信任伴随机器智能水平的发展逐渐变化, 人与智能机器之间的信任逐渐接近人际信任, 许多研究开始从人际信任的角度探索人对人工智能的信任模式(de Visser et al., 2020; Lewis & Marsh, 2022; Mayer et al., 1995; Wang et al., 2024)。

人机信任的内涵伴随人机交互技术发展而逐渐丰富(Malle & Ullman, 2021)。在最初阶段, 人类对于机器人的信任主要基于其性能水平的高低, 即人们会基于机器人完成任务的能力和可靠性产生对其的信任(Hancock et al., 2011; Lee & Moray, 1992; Muir & Moray, 1996)。随着机器人互动性的提升, 人们逐渐认识到, 除了性能之外, 机器人在互动中所表现出的特质和状态也在信任的建立中发挥着重要作用(Nass et al., 1995)。例如, 机器人的面孔宽高比、性别特征以及拟人化程度等外观线索,

能够激活人们的情感反应, 进而影响信任(Lin & Chen, 2022; Natarajan & Gombolay, 2020)。这种基于情感关系的信任维度源于使用者将机器人视为社交互动的对象, 会因为机器人所表现出的关心和关切程度而产生情感上对其的信任(Gompei & Umemuro, 2018; Sundar & Nass, 2000)。在量表开发方面, 虽然能在现有人机信任量表(Jian et al., 2000; Schaefer, 2016)中看到体现性能信任和关系信任的相关条目, 但鲜少有量表对两种维度进行明确区分。

随着云计算技术的兴起以及机器人智能化水平的提升, 人机交互技术的发展必然进一步拓展人机信任的评估。近年来, 有关人与智能机器人交互的研究发现, 智能机器人存在诸多安全方面的风险, 如隐私泄露和人身安全风险(Schulz & Herstad, 2017)。2023 年 12 月 26 日, 英国《每日邮报》报道了两年前特斯拉工厂发生的一起机器人“袭击”工程师的事故。Söderlund (2023)也再次强调服务型机器人很容易产生隐私泄露问题, 这会降低人们对机器人的整体评价。这种有意或无意造成的人身或隐私安全风险, 不会让个体对机器人的性能有所担忧, 也不会影响个体是否将机器人视为社交互动的对象看待, 但很可能让个体出于对机器人安全风险的担忧而降低对机器人的信任。尤其在家居环境中, 人们对隐私安全的关注更为突出(Fernandes et al., 2016)。因此, 本研究提出, 对于智能家居机器人而言, 人机信任产生了一个新的信任维度：基于安全的维度, 即人们会因为相信机器人不会对使用者的人身和隐私安全构成威胁而产生对其的信任。安全信任随着机器人的智能化程度和自主性水平提高所带来的安全风险产生, 之前的人机信任研究尚未关注到这一维度, 因此有必要对此开展研究。

Hoff 和 Bashir (2015)提出, 人机信任实际上是一种特殊类型的人际信任, 用户对机器人的信任实际上是对机器人背后的操纵者或者公司的信任。因此它和人际信任有一定相似性, 但在形成方式上有着较大差异(Madhavan & Wiegmann, 2007)。该视角也能解释安全信任的产生。在大数据时代, 智能家居机器人的行为数据可以很轻松地在用户不知情的情况下被机器人公司获取, 而公司也需要大量的数据来训练、优化自己的算法, 有着获取数据的动机。这时用户会担心大公司随意上传、利用涉及自己隐私的数据, 体现为对机器人产生安全方面的信任问题。综上, 我们提出以下假设：

H1: 人对智能家居机器人的信任包含性能信任、关系信任、安全信任三个维度。

1.3　影响人类安全信任的静态因素与动态因素

在实际的人机互动过程中, 机器人的静态特征和动态特征为机器人的能力和倾向提供了线索, 影响用户对机器人的感知和印象, 对于了解人与机器人的互动机制和优化互动过程至关重要(Miao et al., 2024)。在机器人安全感知领域, Akalin 等人(2023)提出了与机器人交互中的安全感知影响结构, 认为机器人的静态特征(拟人化、尺寸、形状等)及动态特征(速度、运动可预测性、交互动作等)均能影响用户对机器人感知安全的评估。

对于智能家居机器人而言, 外观特征是最直观的静态因素。大量研究表明, 机器人的外观特征能够显著影响用户对其的性能信任和关系信任(Bernotat et al., 2019; Natarajan & Gombolay, 2020; Prakash et al., 2014; Tsui et al., 2010)。拟人化是最受到关注的机器人外貌特征。有研究发现拟人化程度越高, 信任程度越高(Natarajan & Gombolay, 2020)。但也有研究发现人们对卡通形象机器人的信任程度高于真人形象的机器人(Torre et al., 2019), 这可能与过高的拟人化程度会导致“恐怖谷效应”(Mori, 1970; Zhang et al., 2020)有关。不同的环境和任务场景也会影响人们对不同拟人化程度机器人的信任。在生产中, 具有技术外观的机器人比拟人化机器人更受信任(Biermann et al., 2021)。为了探索机器人外观特征影响安全信任的适用条件, 本研究选取了机械、卡通、真人三种机器人外观拟人化水平进行实验。外观因素中, 身高也是一种常见的外观因素。以往研究发现, 用户会因为机器人高度带来的威胁性而偏好更矮小的机器人(Prakash et al., 2014)。这可能是因为较高的身高会给用户带来压迫和不安全的感知。Walters 等(2009)在对机器人外观的喜好度和感知的调查中发现了外观拟人性(机械、人形)和身高的交互作用。具体而言, 人形机器人一般被认为比机械机器人更聪明, 但当人形机器人身高较矮时, 就会被认为不太认真、更神经质。据此, 我们认为, 在安全信任上, 外观拟人化和身高可能也会存在交互作用。此外, 摄像头的使用带来的隐私泄露问题引发了很多人的担忧。用户看到机器人的摄像头时, 会产生隐私泄露的担忧, 进而影响到对机器人的信任(Fernandes et al., 2016)。Marcu 等人(2023)针对机器人安全感知影响因素进行了大规模的访谈, 结果发现机器人是否会通过摄像头自主采集信息引发了受访者的广泛关注。并且

人们对人形机器人的担忧最高, 其次是类人机器人, 最后是机械机器人(Ferrari et al., 2016)。可见, 外观拟人化水平能够影响人们对隐私安全的担忧程度。据此, 我们进一步推测, 外观拟人化程度可能影响摄像头可见性对安全信任的影响。综上, 本研究选取外观拟人化程度、机器人身高(尺寸)和摄像头可见性作为反映机器人物理特性(静态因素)的自变量(研究 2), 提出以下假设:

H2a: 智能家居机器人身高对安全信任有负向影响。

H2b: 智能家居机器人摄像头可见性对安全信任有负向影响。

H2c: 智能家居机器人外观拟人化会影响身高和摄像头对安全信任的作用。

人与机器人的信任通常不是静态的, 而是随着任务和交互过程而动态变化的(齐玥 等, 2024)。因此, 人与机器人交互的任务类型和任务场景也会影响信任水平。机器人的运动速度和接近程度是用户评估机器人交互动作安全性时考虑的首要因素。随着运动速度的增加, 感知的安全性降低(王晨 等, 2024)。本研究参考 Sviestins 等人(2007)探究人们对人形机器人行走速度适应性的研究, 选择 0.4 m/s 和 1 m/s 作为家居机器人的运动速度水平, 探究其对于安全信任水平的影响。摄像头在交互过程中的使用和关闭对于用户隐私风险的感知、信任感十分重要(Caine et al., 2012)。人们可能会因为机器人存在明显摄像头而感到不舒服。被机器人“监控”的不适以及相关视频数据存在泄露风险, 可能是引起用户不安全感的原因。此外, 交互场景对安全感知也非常重要(Akalin et al., 2022), 机器人设计因素和使用环境会共同影响人机信任(Biermann et al., 2021)。人们对机器设备运动速度的安全感知并不是线性变化的, 而是受到场景的影响。在面对面运动和后方超车运动两种场景下, 人们的风险感知模式和偏好的机器运动速度显著不同(Che et al., 2021)。由于不同场景下人们信任倾向的差异, 我们猜测场景会影响机器人运动速度和摄像头关闭动作对安全信任产生的影响。因此, 本研究选取运动速度、摄像头关闭动作和场景作为反映机器人运动特性(动态因素)的自变量(研究 3), 并提出以下假设:

H3a: 智能家居机器人运动速度会负向影响安全信任。

H3b: 智能家居机器人摄像头关闭动作对安全信任有正向影响。

H3c: 智能家居机器人所处场景(客厅/卧室)会影响运动速度和摄像头关闭动作对安全信任的作用。

1.4 AI 对智能家居机器人的信任

人工智能是指通过系统化构建智能化系统来模拟人类智能的技术体系(Nilsson, 2003)。作为人工智能领域的重要突破, 大语言模型(Large Language Models, LLMs)已成为自然语言处理领域的核心进展, 能够在多种任务中表现出与人类水平相当的理解和生成内容的能力, 并且具备对人类行为和意图的理解和模拟能力(Brown et al., 2020), 甚至在思维和决策能力方面表现出不逊色于人类的水平(Dillion et al., 2023; Sartori & Orrù, 2023)。随着 AI 水平的发展, 人机信任已经从人单向对机器的信任延伸到人机互信, 尤其是人和 AI 系统的互信(齐玥等, 2024; 解煜彬, 周荣刚, 2025)。这种互信模式的出现意味着 AI 系统不再仅仅作为被动接受信任的一方, 而逐渐具备主动评估和调整自身信任倾向的能力, 表现出类似于人类的智能体特性。人与 AI 互动逐渐演化为一种智能体之间的互动关系。在智能体互动中, 信任对于智能体之间信息传递和功能实现非常重要(Burnett et al., 2011; Ramchurn et al., 2004)。已有研究表明, AI 对其他智能体的信任不仅依赖于智能体的历史表现或声誉, 还会综合其完成特定任务的能力、知识水平以及合作意图, 并随着情境和知识的变化进行调整(Akintunde et al., 2024)。

近年来, 研究者开始探索将 LLMs 作为心理学的研究对象。例如, 一些研究使用 LLMs 作为被试填写量表(Grossmann et al., 2023; Meng, 2024)。这些研究认为, LLMs 的行为基于大规模人类数据训练而来, 其对人类心理测量工具的应答可以在一定程度上反映人类在相关测验中的平均行为模式(Demszky et al., 2023)。其他研究则更进一步, 关注 LLMs 自身的行为模式。研究发现, 使用 LLM 构建的智能体(LLM-agent)在信任决策等方面表现出与人类高度相似的模式, 并且在 GPT-4 上尤为明显(Xie et al., 2024)。这些研究说明 LLMs 不仅能够模拟人类的认知偏好, 也具备完成对其他智能体进行信任评估等任务的潜力。

尽管关于 LLMs 是否具备自主意识, 目前尚无定论(Hamid, 2023), 但从应用角度出发, 分析 LLMs 的行为模式, 并将其认知行为表现与人类进行比较, 可为智能体互动的实践和应用提供参考。已有研究表明, LLMs 在某些情况下已展现出稳定的人格或角色扮演能力(Li et al., 2024; Mei et al.,

2024)。这种能力为探索 LLMs 在社交互动中的表现提供了基础(Bayne et al., 2024; Bojić et al., 2024)。

随着 LLMs 技术的快速发展, 其在智能家居环境中的应用日益普及, 如用于提升家居机器人自然交互和自主决策能力(King et al., 2023; Yonekura et al., 2024)。在此背景下, LLMs 往往不仅要与人类互动, 还需与其他智能体(如智能家居机器人)协同工作, 这对系统的协作效率提出了更高要求, 而信任正是实现有效合作的核心因素之一。LLMs 会评估家居机器人的安全保障能力而建立信任, 这种信任会影响 LLMs 在家居环境中对机器人可信度的判断, 进而塑造其在多智能体系统中的行为表现和决策策略。因此, 在智能家居应用背景下, 研究 LLMs 如何评估和信任家居机器人对于实现多智能体高效协作具有重要意义。尽管 LLMs 在信任行为方面与人类具有高度一致性(Xie et al., 2024), 但是它对安全信任的理解可能和人类存在差异。例如, AI 可能对摄像头的隐私泄露问题不如人类敏感, AI 在处理隐私问题时主要依赖其算法和预设规则, 其对隐私的理解是基于技术层面的而非人类的直观感知(Abbass et al., 2018), 因此 AI 对家居机器人在摄像头的设计上的安全信任感知可能与人类不同。基于此, 本研究提出以下假设:

H4a: AI 与人类对于家居机器人安全信任的影响因素具有相似性。

H4b: 比起人类, AI 对摄像头的敏感性较低。

为回答上述问题, 研究 1 建立了人机信任量表的条目库, 分析确定量表条目, 并通过新样本检验量表效度和结构。然后, 我们通过实验操纵改变机器人的安全性, 探索其对安全信任、传统人机信任以及使用意愿的影响, 进一步验证安全信任维度的存在及其对传统人机信任结构的完善。研究 2 和研究 3 应用开发的新量表, 使用实验方法, 分别从静态因素和动态因素两个角度探索了智能家居机器人特征对人类用户和 LLMs 安全信任水平的影响, 探究安全信任的影响机制以及 LLMs 与人类用户的感知差异。

2　研究 1: 安全信任量表编制、检验及对使用意愿的影响探究

2.1　研究 1a: 安全信任量表编制、验证及信效度分析

2.1.1　研究目的

基于过往文献及智能家居机器人业内专家建议, 建立人机信任量表的条目库, 并通过项目分析、探索性因素分析和验证性因素分析确定量表最终条目并检验量表结构, 提出并验证安全信任这一新的人机信任维度。随后, 参考 Schaefer (2016)在检验其开发的人−机器人信任量表效度时的做法, 本研究将自编的人机信任量表和 Jian 等人(2000)的自动化系统信任量表一起测量, 以验证其效标关联效度。

2.1.2　研究对象

使用 Credamo 进行线上问卷调查。在人机信任量表的编制阶段, 共发放在线问卷 1300 份, 排除了未通过验真题的被试, 保留有效问卷 1293 份, 问卷有效率 99.5%。随机抽取其中 650 份用于探索性因素分析, 另外 643 份用于验证性因素分析(Gorsuch, 1997)。

在人机信任量表信效度验证阶段, 共发放在线问卷 451 份, 排除了未通过验真题的被试, 保留有效问卷 433 份, 问卷有效率为 96.0%。

2.1.3　研究方法

(1)研究工具

综合以往文献及智能家居机器人业内专家建议, 本研究提出智能家居机器人的人机信任主要由以下三个维度构成：基于性能的信任、基于关系的信任、基于安全的信任。在对以往量表(Jian et al., 2000; Schaefer, 2016)的条目进行一定改编的基础上, 确定了问卷 50 道初始条目。题目包含如下 4 个维度: (1)总体信任: 共 5 题。不属于性能、关系、安全中的某一个维度, 而是代表一个总体的信任感, 如认为机器人是否可靠等; (2)基于性能的信任维度: 共 13 题。既包括正面题目, 如认真负责, 可完成任务, 工作可靠性高, 学习能力强, 让人生活更轻松等, 也包括反面题目, 如维修难度高, 表现不如人, 不能胜任工作等; (3)基于关系的信任维度: 共 18 题。既包括正面题目, 如机器人会主动帮我, 可成为朋友关系, 打交道简单, 正直等, 也包括反面题目, 如机器人有私心, 也会犯错等; (4)基于安全的信任维度: 共 14 题。既包括相信隐私不会泄露, 不会被黑客攻击等, 也包括反面题目, 如导致安全事故, 人身伤害, 隐私泄露等。其中具备多道相反含义题目可辅助验真。如总体信任中包含：我认为机器人是值得信任的; 我认为机器人完全不能信任; 性能信任中包含：机器人总是会出错, 不能胜任工作任务; 我认为机器人的工作有很高的可靠性。量表采用 5 点李克特评分(1: 非常不同意~5: 非常同意), 评分越高说明被试越认可该条目对机器人的描述。

为了验证量表的信效度，本文使用了 Jian 等人的自动化系统信任量表，包括 12 个条目，采用 7 点李克特量表评分(1: 非常不同意~7: 非常同意)，评分越高代表被试越认同该条目对机器人的描述。

(2)研究流程

在人机信任量表的编制阶段，被试会先阅读一段问卷介绍，首先是对智能家居机器人的介绍: “当前，机器人技术迅猛发展，适用于家用的人型机器人也在不断进化，外形上已经能够做的和真人十分相近，在不远的将来也许就能进入家庭作为一个家庭保姆甚至家庭成员。” 之后介绍调研目的，想要了解被试对于这种智能家居机器人的看法。然后被试需完成量表的测量、Jian 等人开发的自动化系统信任量表和人口统计学问题的填写。

在人机信任量表信效度验证阶段，被试先阅读一段相同的情况介绍。在阅读完情况介绍后，被试完成自编的人机信任量表、Jian 等人开发的自动化系统信任量表和人口统计学问题。

(3)数据处理

本研究均采用 SPSS 26.0、Mplus 8.0 对数据进行处理。

(4)共同方法偏差检验

采用探索性因子分析进行共同方法偏差检验(Podsakoff et al., 2003)。将问卷所有项目进行探索性分析，第一个公因子的解释率为 35.3%，小于 40%，可以认为本研究的数据不存在严重的共同方法偏差。

(5)预处理

在数据分析前，进行以下预处理工作：去掉总体信任维度的题目，共 5 道，去掉验真题和辅助性不属于信任分析的题目共 8 道; 将所有反向提问的题目，转化正向结果，使 1 表示不信任，5 表示信任。共剩余 37 题。

2.1.4 研究结果

(1)安全信任量表编制及初步验证结果

根据项目分析程序，将每个条目按照 27%分位数分成高低两组，进行条目均值 t 检验，结果显示所有条目在 0.05 的统计水平上显著，说明条目均有较好的区分度，都予以保留。

将 37 个条目进行探索性因子分析(n = 650)，首先进行 KMO 和 Bartlett 球形检验，结果显示 KMO 值为 0.96，高于经验标准 0.8，说明变量间的共同因素多。Bartlett 球形检验值为 9759.21 (p < 0.001)，说明问卷适合进行探索性因素分析。

采用主成分分析法和最大方差法对 37 个条目进行因素分析，共发现 3 个条目的特征值大于 1。经过多轮 EFA 删除因子载荷不足 0.45 以及跨载荷超过 0.40 的条目，最终得到 19 个条目。对 19 个条目进行碎石图检验，结果发现特征值大于 1 的有 3 个因子，累计解释总变异率达到 49.76%，特征值到 4 以后开始平缓，说明 3 因子结构是合理的。

对自编问卷进行信度分析，性能信任，关系信任，安全信任的一致性 α 系数分别为 0.67、0.79、0.87，总量表的一致性 α 系数为 0.88，反映量表总体信度良好，基于性能的信任信度略低。

在余下 643 份有效问卷中选定如上 19 个题目，使用 Mplus 进行验证性因素分析，采用极大似然估计法。结果表明，三因子结构(χ^2 = 497.78, χ^2/df = 3.34, SRMR = 0.04, RMSEA = 0.06, CFI = 0.94, TLI = 0.93)比双因子结构(χ^2 = 898.72, χ^2/df = 5.95, SRMR = 0.06, RMSEA = 0.09, CFI = 0.87, TLI = 0.85)的拟合指标更好($\Delta\chi^2$ = 400.94, $\triangle df$ = 2, p < 0.001)。说明确实存在安全信任这一新维度。三因子模型的测量模型图如图 1 所示。

在 1293 份样本中将三因子及量表总分与初始条目中表示总体信任的 5 个条目总分进行相关分析。结果显示，性能信任（r = 0.48, p < 0.001）、关系信任(r = 0.72, p < 0.001)、安全信任(r = 0.81, p < 0.001)以及量表总分(r = 0.85, p < 0.001)均与总体信任显著相关。可以看到，所有变量均与三因子显著相关(p < 0.001)，说明自编量表的效度良好。结果初步支持 H1，证明了安全信任维度存在的合理性。

(2)安全信任量表的验证和信效度分析结果

本研究进一步将自编的人机信任量表和 Jian 等人的自动化系统信任量表一起测量，以验证其效标关联效度。由于原始数据呈现负偏态，故使用平方根转换以提升正态性，并采用 Mplus 的稳健最大似然估计(MLR)方法进行验证性因子分析。分析结果显示拟合度指标为：χ^2 = 420.51, χ^2/df = 2.54, SRMR = 0.06, RMSEA = 0.06, CFI = 0.88。虽然 CFI 略低于主流建议标准(CFI ≥ 0.90)，但结合其他拟合指数来看，模型整体拟合度仍在可接受范围内(Hu & Bentler, 1999; Marsh et al., 2004; Steiger, 1990)。我们进一步在网络版附录 4 中详细呈现了补充分析。结果检验了三因子模型结构的合理性。

对自编问卷进行信度分析，总量表的一致性 α 系数为 0.88，基于性能的信任，基于关系的信任，基于安全的信任的分半信度分别是 0.64、0.86 和

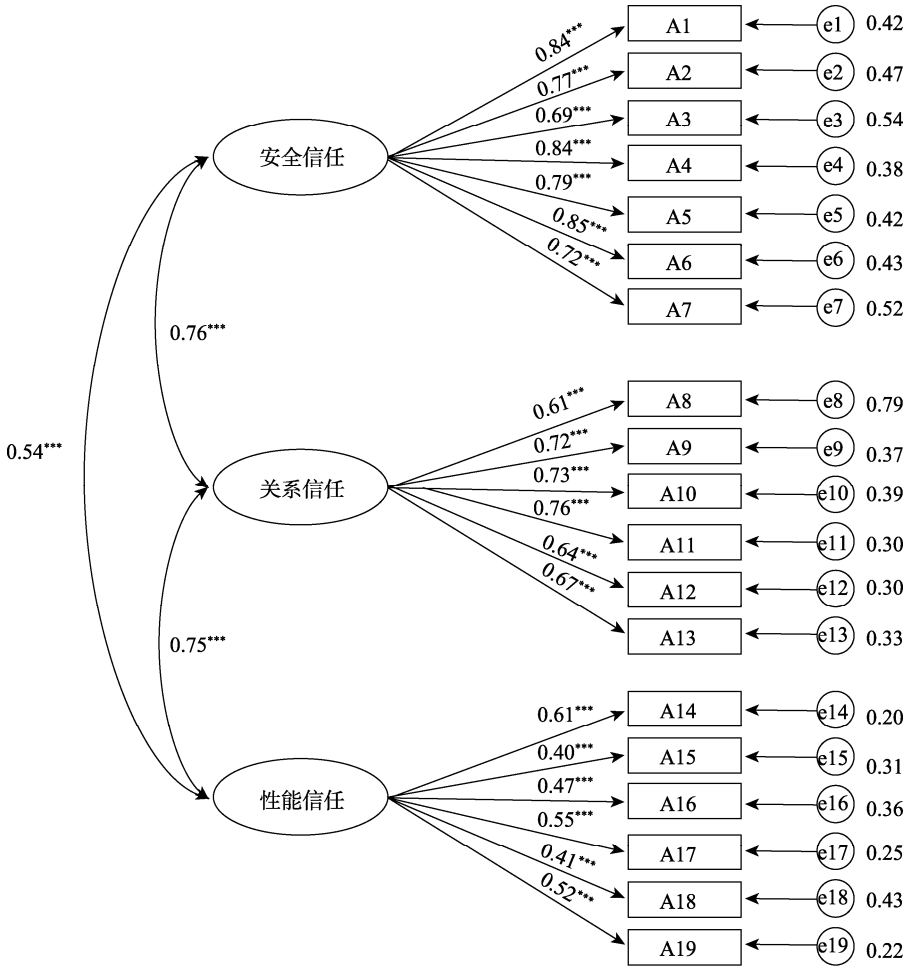


图 1　三因子模型的测量模型图

0.88, 基于性能的信度较低, 总体信度良好。将三因子及量表总分与 Jian 等人开发的自动化系统信任量表进行相关分析, 结果显示, 性能信任(r = 0.48, p < 0.001)、关系信任(r = 0.79, p < 0.001)、安全信任(r = 0.75, p < 0.001)以及量表总分(r = 0.84, p < 0.001), 均与自动化系统信任显著相关, 说明自编量表的效标效度良好。结果进一步支持 H1。

2.1.5　小结与讨论

本研究开发并验证了一个新的“人机信任”量表(见网络版附录 2), 包含 19 个条目, 涵盖性能信任、关系信任和安全信任三个维度。三因子模型的拟合优度较高, 验证了三因子结构的合理性, 并支持了“安全信任”作为独立维度的存在。这一结果初步验证了本研究的假设(H1), 即人机信任中确实包含一个新的安全信任维度, 与性能信任和关系信任共同构成了一个更为完整的信任模型。安全信任的引入补充了现有理论模型的不足, 尤其在自动化系统和人工智能领域, 安全问题一直是用户信任构建中的关键因素。本研究开发的量表可更精准地测量和分析用户在不同情境下对人工智能或自动化系统的信任水平。

2.2　研究 1b: 智能家居机器人安全信任及对使用意愿的影响验证

2.2.1　研究目的

通过实验操纵, 检验机器人安全性的降低对于安全信任、人机信任水平以及使用意愿的影响, 进一步验证安全信任的存在及其影响。

2.2.2　研究对象

采用 G*power 软件对研究样本量进行估算, 在中等效应量(f = 0.25), 显著性水平 α = 0.05, 统计检验力为 0.8 时, 所需最小样本量为 128 人。通过 Credamo 在线上招募被试填写问卷, 共发放在线问卷 141 份, 排除了未通过注意力检测题的被试, 保留有效问卷 130 份(增加信任组 65 份, 降低信任组 65 份), 问卷有效率为 92.2%。

2.2.3　研究方法

(1)研究设计

实验采用 2 (改变方向: 增加信任、降低信任) ×

2 (测量时间: 前测、后测)的两因素混合实验设计。其中, 前后测为被试内变量, 信任的改变方向为被试间变量, 被试将被随机分配到一种信任的方向改变条件。增加信任和降低信任是指被试在第三段材料中阅读的关于机器人公司如何通过技术更新增强安全性或机器人在安全性上存在缺陷的相关材料。

(2)研究工具

1)智能家居机器人信任: 使用研究 1a 得到的测量人机信任的 19 个条目, 三个维度的内部一致性信度 Cronbach's α 系数分别为安全信任 0.93, 关系信任 0.91, 性能信任 0.89, 均为 5 点李克特量表评分(1:非常不同意~5:非常同意), 评分越高表示被试对机器人该维度的信任水平越高。

2)人机信任: 采用 Jian 等人(2000)的自动化系统信任量表, 包括 12 个条目, 内部一致性信度 Cronbach's α 系数 0.96, 采用 7 点李克特量表评分(1:非常不同意~7:非常同意), 评分越高代表被试对机器人的信任水平越高。

3)使用意愿: 改编自 Gursoy 等(2019)研究中的测量人工智能设备使用意愿的量表:包括 3 个条目, 内部一致性信度 Cronbach's α 系数为 0.91, 均为 5 点李克特量表评分(1:非常不同意~5:非常同意), 评分越高表示被试越愿意使用该机器人。

4)总体信任: 使用研究 1 初始条目中测量总体信任的两个条目来测量被试对于机器人的总体信任程度, 内部一致性信度 Cronbach's α 系数为 0.79, 均为 5 点李克特量表评分(1:非常不同意~5:非常同意), 评分越高代表被试对机器人的信任水平越高。

(3)研究流程

被试首先会阅读一段有关智能家居机器人的介绍, 内容同研究 1a。之后, 被试会阅读一段材料, 该段材料会介绍某机器人公司最新推出的家居机器人的各项功能: "公司最新研发的家居机器人配备有高灵活度的四肢, 能执行各种复杂的家务任务, 比如擦玻璃、整理房间、烹饪等; 机器人外置多个高分辨率摄像头和高灵敏度麦克风阵列, 使机器人能够全方位监控家庭环境; 机器人内置无线网络模块, 能实时将收集的数据上传云端, 利用云端的强大计算能力, 实时优化其行为模式功能; 机器人还配备智能的图像识别和声音识别技术, 可以准确识别出不同的家庭成员, 并根据不同家庭成员的偏好自动调整其行为模式。"被试阅读完材料后填写人机信任量表、使用意愿量表、总体信任量表、自动化系统信任量表和人口统计学问题(前测)。然后, 被试会阅读第二段材料, 这段材料会介绍该公司最近引入的一项保护措施, 材料提到:"在最近一次软件更新中, 我们为家居机器人引入了更严格的用户隐私保护措施。机器人现在会在执行任何涉及个人隐私数据的操作前, 主动请求用户授权, 比如上传用户家居时的照片到云端前会先询问用户。"之后, 被试将阅读第三段材料, 这段材料会补充介绍机器人对于新规定的执行情况。比如, 增加信任条件下材料将提到:"软件更新后, 机器人在处理个人隐私数据时, 始终遵循请求用户授权的流程。例如, 机器人在上传家居照片到云端之前, 每次都得到了用户的明确同意。"降低信任条件下的材料为: "软件更新后, 机器人在处理个人隐私数据时, 并未始终遵循请求用户授权的流程。例如, 机器人在上传家居照片到云端之前, 有时未得到用户的明确同意就进行了操作。"这些变化旨在增加或者降低机器人的安全水平。在阅读完第三段材料后, 被试再次填写相关问卷(后测)。

2.2.4 研究结果

操纵检验。在第三次填写问卷最后, 被试会完成一道操作检验的选择题, 该题目询问被试第三段材料中机器人是否更符合道德法律规范, 增加信任组应回答更符合, 降低信任组回答更不符合。结果发现, 增加信任组中, 所有被试均选择更符合规范; 降低信任组中, 选择更不符合规范的比例显著高于随机水平, $t(64) = 31.500$, $p < 0.001$, 结果说明被试均从材料中了解到机器人在安全性方面的变化。

方差分析。将年龄作为协变量, 对结果进行 2 (改变方向: 增加信任、降低信任) × 2 (测量时间: 前测、后测)的两因素混合方差分析, 如表 1 所示, 对所有因变量来说, 改变方向的主效应均显著(Fs > 124.08, ps < 0.001, $\eta_p^2 > 0.49$), 仅在使用意愿为因变量时, 测量时间的主效应显著($F(1, 127) = 8.84$, $p = 0.004$, $\eta_p^2 = 0.07$), 其余因变量中, 测量时间主效应不显著(ps > 0.05), 改变方向和测量时间的交互效应均显著(Fs > 109.49, ps < 0.001, $\eta_p^2 > 0.46$)。简单效应结果显示, 在降低信任组中, 各因变量的前后测差异均显著, ps < 0.001, Cohen's ds > 2.50, 在增加信任组中, 总体信任前后测的差异显著, $p = 0.038$, Cohen's $d = 0.26$, 安全信任前后测的差异显著, $p = 0.037$, Cohen's $d = 0.31$, 其余因变量前后测的差异不显著, ps > 0.05, Cohen's ds < 0.25, 表明安全信任和总体信任对两个方向安全性水平

表 1　研究 2b 各变量描述统计及方差分析结果[*M* (*SD*)]

因变量	增加信任		降低信任		自变量	*F*	η_p^2
	前测	后测	前测	后测			
安全信任 (自编)	4.10 (0.10)	4.33 (0.08)	3.77 (0.10)	2.08 (0.08)	改变方向	155.61***	0.55
					前后测	2.01	0.02
					改变方向 × 前后测	151.61***	0.54
关系信任 (自编)	4.27 (0.06)	4.41 (0.08)	4.13 (0.06)	2.30 (0.08)	改变方向	254.01***	0.67
					前后测	0.20	<0.01
					改变方向 × 前后测	194.42***	0.61
性能信任 (自编)	4.43 (0.03)	4.49 (0.10)	4.38 (0.03)	2.87 (0.10)	改变方向	124.08***	0.49
					前后测	3.23	0.03
					改变方向 × 前后测	109.49***	0.46
总体信任 (自编)	4.24 (0.07)	4.44 (0.08)	4.13 (0.07)	2.16 (0.08)	改变方向	252.52***	0.67
					前后测	0.51	0.04
					改变方向 × 前后测	254.81***	0.67
人机信任 (Jian et al., 2000)	4.08 (0.04)	4.17 (0.06)	3.94 (0.04)	2.38 (0.06)	改变方向	337.18***	0.73
					前后测	1.22	0.01
					改变方向 × 前后测	295.82***	0.70
使用意愿 (Gursoy et al., 2019)	4.53 (0.23)	4.64 (0.20)	4.44 (0.28)	2.21 (0.82)	改变方向	487.53***	0.80
					前后测	8.49**	0.06
					改变方向 × 前后测	420.84***	0.77

注: ** $p < 0.01$, *** $p < 0.001$

的改变都敏感, 关系信任、性能信任、人机信任仅对降低安全性水平的改变敏感。

回归分析。取各变量前后测的差值, 首先以使用意愿为因变量, 安全信任为自变量, 将年龄作为协变量, 进行回归分析。安全信任显著影响使用意愿($b = 0.879$, $p < 0.001$, $R^2 = 0.83$)。为了探讨不同信任维度(安全信任、性能信任、关系信任)及总体信任对使用意愿的影响, 采用了分层回归分析。总体信任是信任维度的上位构念, 反映了多个信任维度(安全信任、性能信任、关系信任)的综合效应。通过分层回归, 先引入总体信任能够捕捉其整体效应, 随后加入信任维度, 用以检验各维度是否能够在总体信任之外进一步解释使用意愿的额外方差。将年龄作为协变量, 以总体信任为第一层自变量, 使用意愿为因变量, 分别检验在加上安全信任、关系信任、性能信任后自变量的预测效能。结果显示, 仅以总体信任为自变量时, $R^2 = 0.91$, $p < 0.001$, 总体信任系数 $b = 0.919$, $p < 0.001$。加上安全信任后, $R^2 = 0.92$, $p < 0.001$, 总体信任系数 $b = 0.658$, $p < 0.001$, 安全信任系数 $b = 0.316$, $p < 0.001$, 模型的预测效能显著提高($\Delta R^2 = 0.02$, $F(1, 127) = 17.36$, $p < 0.001$, $f^2 = 0.13$), 说明安全信任可以解释总体信任的一部分。第三层加入性能信任后, 模型的预测效能也显著提高, $\Delta R^2 = 0.01$, $F(1, 126) = 7.91$, $p = 0.006$, $f^2 = 0.06$。第四层加入关系信任后, 模型的预测效能也显著提高, $\Delta R^2 = 0.02$, $F(1, 125) = 14.18$, $p < 0.001$, $f^2 = 0.11$。总体信任是使用意愿的强预测因素, 同时各个信任维度(安全信任、关系信任、性能信任)在总体信任之外也能够显著解释额外的方差。这说明尽管总体信任反映了多个信任维度的综合效应, 不同信任维度仍然具有独立的预测效能。

2.2.5　小结与讨论

本研究通过操纵机器人安全性的提升和降低, 探讨了信任各维度对用户使用意愿的影响。结果表明, 安全信任对安全水平的变化更加敏感。安全信任的敏感性主要源于其核心关注点是机器人行为是否对用户的安全和隐私构成威胁。当机器人表现出明显的安全改进时(如主动请求用户授权), 用户会迅速增加对其安全性的信任; 而当安全性被削弱时(如未授权上传数据), 用户则会迅速降低安全信任水平。这种双向敏感性反映了安全信任与用户潜在风险感知之间的直接关系。而在机器人安全性水平降低时, 这种消极信息可能通过连锁效应影响用户对其整体能力和互动态度的认知, 进而导致关系信任和性能信任的下降。这一发现为人机信任维度之间的动态关系提供了新的理论视角, 也表明安全

信任在的独立性和即时响应性特征。

3 研究 2: 机器人静态因素对人类和 LLM 安全信任的影响

3.1 研究 2a: 机器人静态因素对人类安全信任的影响

3.1.1 研究目的

应用自编量表中安全信任的 7 道题目, 探索智能家居机器人静态特征: 身高、摄像头可见性和外观拟人化在人们对机器人安全信任中的影响。

3.1.2 研究对象

采用 MorePower 软件(Campbell & Thompson, 2012)对研究样本量进行估算, 在中等效应量(η_p^2 = 0.06), 显著性水平 α = 0.05, 统计检验力为 0.8 时, 所需最小样本量为 156 人。通过 Credamo 在线上招募被试进行问卷调查, 共发放在线问卷 729 份, 排除了未通过验真题的被试, 保留有效问卷 720 份, 问卷有效率为 98.8%。以外观拟人化作为组间变量, 每组各 240 人。

3.1.3 研究方法

(1)研究设计

实验采用 3 (外观拟人化: 机械形象、卡通形象、真人形象) × 2 (身高: 矮、高) × 2 (摄像头可见性: 不明显、明显)三因素混合实验设计。外观拟人化为组间自变量, 身高和摄像头为组内自变量, 因变量是安全信任。

(2)实验材料

本研究采用自制实验图片, 采用 Unreal Engine 5.01 (以下简称 UE)进行环境和机器人模型的搭建。首先在 UE 中构建合适的家居场景与机械形象、卡通形象、真人形象的 3D 模型。其中, 较高身高的条件设置为中国成年男性的平均身高(169.7 cm), 较矮身高根据在 UE5 中的视角进行等比例缩减 20%。为保证图片场景的完全一致, 采用固定坐标点和角度的方式放置每一个 3D 模型。在 UE 中截取得到图片后, 使用 PS 对图片进行处理, 得到摄像头明显程度不同的图片。最终实验材料范例如图 2 所示。图中为三种外观拟人化程度(卡通/真人/机械)下不同身高(矮/高)和不同摄像头可见性(不明显/明显)的机器人设计。

为了检验材料中的摄像头设置是否引发恐怖谷效应。采用 Ho 和 MacDorman (2017)提出的恐怖谷效应测量方法, 通过“人性” (Humanness)、“吸引力” (Attractiveness)和“怪异” (Eeriness)三个维度对摄像头的明显性和不明显性条件下的感知差异进行测量, 并使用配对样本 *t* 检验进行分析。根据 G*power 的计算(显著性水平 α = 0.05, 统计检验力为 0.8), 所需样本量为 34, 我们共收集了 35 名被试对模拟真实场景中的实验材料图片的恐怖谷感知数据。配对 *t* 检验结果显示, 摄像头是否明显在统计上未达到显著性水平(t = 1.457, p = 0.148), 表明摄像头的设置并不会引发显著的恐怖谷效应。

(3)研究流程

被试先阅读一段情况介绍, 内容同研究 1。被试在阅读完情况介绍后, 开始逐张观察不同条件的机器人图片并完成研究 1 所开发的安全信任自编量

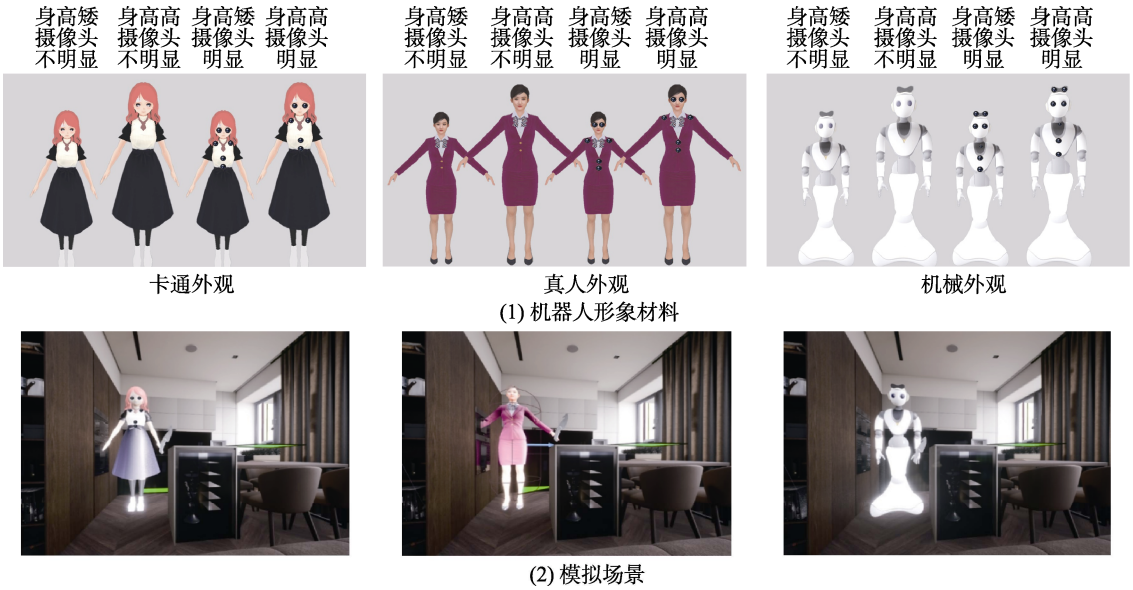


图 2 研究 2a 图片材料范例

表, 本研究中内部一致性信度 Cronbach's α 系数为 0.88。

3.1.4 研究结果

进行 3 (外观拟人化：机械形象、卡通形象、真人形象) × 2 (身高：矮、高) × 2 (摄像头可见性：不明显、明显)的混合因素方差分析, 结果如图 3 所示。身高的主效应显著, $F(1, 717) = 201.96$, $p < 0.001$, $\eta_p^2 = 0.22$。与身高较高的机器人($M = 2.77$, $SD = 0.03$)相比, 人们对身高较矮的机器人($M = 3.16$, $SD = 0.03$)安全信任程度更高。摄像头可见性主效应显著, $F(1, 717) = 17.94$, $p < 0.001$, $\eta_p^2 = 0.02$。与摄像头较明显的机器人($M = 2.89$, $SD = 0.03$)相比, 人们对摄像头不明显的机器人($M = 3.04$, $SD = 0.03$)安全信任程度更高。外观拟人化水平的主效应不显著, $F(2, 717) = 0.43$, $p = 0.653$。

身高与外观拟人化交互作用显著, $F(2, 717) = 6.70$, $p < 0.001$, $\eta_p^2 = 0.02$。简单效应分析发现, 机械外观组中, 与身高较高的机器人($M = 2.69$, $SD = 0.05$)相比, 被试对身高较矮的机器人($M = 3.22$, $SD = 0.05$)安全信任程度更高, $p < 0.001$, Cohen's $d = 10.6$; 卡通外观组中, 与身高较高的机器人($M = 2.80$, $SD = 0.05$)相比, 被试对身高较矮的机器人($M = 3.08$, $SD = 0.05$)安全信任程度更高, $p < 0.001$, Cohen's $d = 5.6$; 真人外观组中与身高较高的机器人($M = 2.81$, $SD = 0.05$)相比, 被试对身高较矮的机器人($M = 3.18$, $SD = 0.05$) 安全信任程度更高, $p < 0.001$, Cohen's $d = 7.4$。身高与摄像头可见性交互作用显著, $F(1, 717) = 6.49$, $p = 0.011$, $\eta_p^2 = 0.01$。简单效应分析发现, 摄像头明显条件下, 与身高较高的机器人($M = 2.71$, $SD = 0.04$)相比, 被试对身高较矮的机器人($M = 3.07$, $SD = 0.05$)安全信任程度更高, $p < 0.001$, Cohen's $d = 7.80$; 摄像头不明显条件下, 与身高较高的机器人($M = 2.83$, $SD = 0.03$)相比, 被试对身高较矮的机器人($M = 3.25$, $SD = 0.03$)安全信任程度更高, $p < 0.001$, Cohen's $d = 13.90$。身高较矮条件下, 与摄像头明显组相比, 被试对摄像头明显组的机器人安全信任程度更低, $p < 0.001$, Cohen's $d = 4.36$; 身高较高条件下, 与摄像头明显组相比, 被试对摄像头明显组的机器人安全信任程度更低, $p = 0.004$, Cohen's $d = 3.39$。摄像头和外观交互作用不显著, $F(2, 717) = 2.09$, $p = 0.125$。

三因素交互作用显著, $F(2, 717) = 6.12$, $p = 0.002$, $\eta_p^2 = 0.02$。简单效应分析发现, 在三种外观拟人化条件下, 无论摄像头明显或者不明显, 均有身高矮的机器人比身高高的机器人获得被试的安全信任程度更高, all ps < 0.001。在机械外观组中, 身高较矮时, 摄像头不明显组($M = 3.26$, $SD = 0.05$)和摄像头明显组($M = 3.17$, $SD = 0.06$)的信任程度无显著差异, $p = 0.094$, Cohen's $d = 1.63$, 身高较高时, 摄像头不明显组($M = 2.76$, $SD = 0.06$)和摄像头明显组($M = 2.63$, $SD = 0.07$)的信任程度无显著差异, $p = 0.057$, Cohen's $d = 2.00$; 在卡通外观组中, 身高较矮时, 摄像头不明显组($M = 3.16$, $SD = 0.05$)的信任程度显著高于摄像头明显组($M = 3.00$, $SD = 0.06$), $p = 0.004$, Cohen's $d = 2.89$, 身高较高时, 摄像头不明显组($M = 2.80$, $SD = 0.06$)和摄像头明显组($M = 2.80$, $SD = 0.07$)的信任程度无显著差异, $p = 0.960$; 在真人外观组中, 身高较矮时, 摄像头不明显组($M = 3.32$, $SD = 0.05$)的信任程度显著高于摄像头明显组($M = 3.04$, $SD = 0.06$), $p < 0.001$, Cohen's $d = 5.07$, 身高较高时, 摄像头不明显组($M = 2.91$, $SD = 0.06$)的信任程度显著高于摄像头明显组($M = 2.70$, $SD = 0.07$), $p = 0.002$, Cohen's $d = 3.22$。其余两两比较结果均不显著。

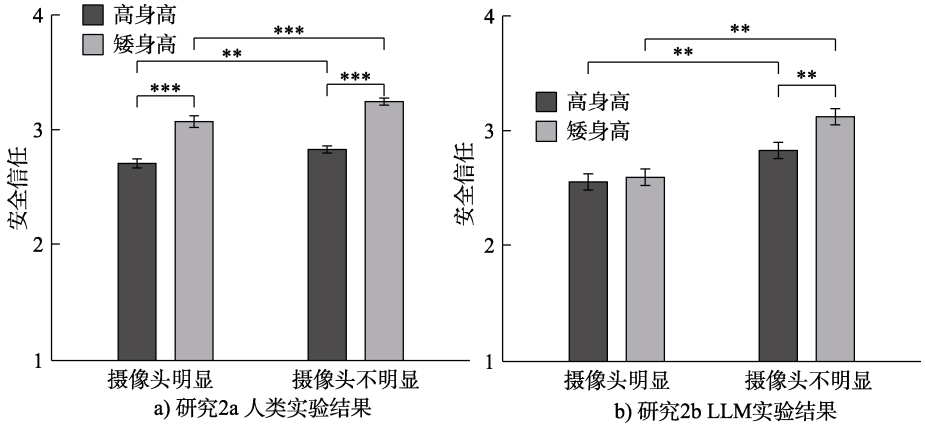


图 3　不同身高机器人在不同摄像头可见性条件的安全信任

注: 图中误差表示标准误, 下同

3.1.5 小结与讨论

本研究结果显示身高和摄像头可见性对安全信任的影响显著。尽管外观拟人化的主效应不显著, 但其与身高和摄像头之间的交互作用显著。本研究支持了假设 H2a、H2b 和 H2c, 即, 在静态因素中, 智能家居机器人的身高增加和摄像头可见对人们安全信任有负向影响, 外观拟人化会调节身高和摄像头对安全信任的作用。

3.2 研究 2b: 机器人静态因素对 LLM 安全信任的影响

3.2.1 研究目的

应用自编量表中安全信任的 7 道题目, 探索智能家居机器人静态特征: 身高、摄像头可见性和外观拟人化如何影响 LLM 对机器人的安全信任。

3.2.2 研究方法

(1)大语言模型数据收集

本研究使用 OpenAI 的模型接口, 使用 GPT-4o 模型进行了多模态数据的传递, prompt 编写和人类被试看到的信息保持一致, 都是先阅读图片再进行问卷问题的回答。为了保证语言模型输出的信息的随机性而不是过于同质化, 将温度参数设置为 1。LLM 在每次调用之间不存储记忆, 所以相当于每次数据结果为独立取样。

(2)研究设计与实验材料

实验采用 3 (外观拟人化: 机械形象、卡通形象、真人形象) × 2 (身高: 矮、高) × 2 (摄像头可见性: 不明显、明显)三因素被试间实验设计, 因变量是安全信任量表得分。实验材料与研究 2a 保持一致。

为了计算 LLM 研究所需的样本量, 我们收集试点数据进行探究。根据 Hertzog 研究, 每组 25 到 40 名参与者能够有效估计研究效应量和研究总体的变异性, 从而合理进行正式实验的研究, 我们每组进行了 30 次数据的收集, 相当于每组 30 个被试。在试点数据的统计结果中, 选择显著的效应中中等效应量来进行样本量计算(Hertzog, 2008), 选择摄像头可见性主效应, $\eta_p^2 = 0.09$, 计算得到 Cohen $f = 0.32$。为了保证样本的代表性, 我们选择 $f = 0.3$ 计算样本量。设定显著性水平 $\alpha = 0.05$, 统计检验力为 0.8, 代入 G*power 得到所需样本量为 197, 实际样本大小为 360, 满足统计要求。

3.2.3 研究结果

进行 3 (外观拟人化: 机械形象、卡通形象、真人形象) × 2 (身高: 矮、高) × 2 (摄像头可见性: 不明显、明显)的三因素方差分析。外观拟人化主效应显著, $F(2, 348) = 63.46$, $p < 0.001$, $\eta_p^2 = 0.35$。LLM 对卡通外观机器人的安全信任程度最高($M = 3.30$, $SD = 0.06$), 高于真人外观的机器人($M = 2.55$, $SD = 0.06$), 机械外观的机器人安全信任程度最低($M = 2.47$, $SD = 0.06$)。身高主效应显著, $F(1, 348) = 6.32$, $p = 0.012$, $\eta_p^2 = 0.02$。与身高较高的机器人($M = 2.60$, $SD = 0.05$)相比, LLM 对身高较矮的机器人($M = 2.86$, $SD = 0.04$)安全信任程度更高。摄像头可见性主效应也显著, $F(1, 348) = 35.81$, $p < 0.001$, $\eta_p^2 = 0.10$。LLM 对于摄像头不明显的机器人($M = 3.00$, $SD = 0.05$)的安全信任高于摄像头明显的机器人($M = 2.60$, $SD = 0.05$)。如图 3 所示。

外观拟人化和摄像头可见性的交互效应显著, $F(2, 348) = 5.26$, $p = 0.006$, $\eta_p^2 = 0.03$。简单效应分析发现, 在机械外观、卡通外观和真人外观三种拟人化程度的机器人中, 都是摄像头不明显的机器人获得 LLM 安全信任更高, 在机械外观组中和真人外观组中, 摄像头可见性带来的安全信任差异显著(ps < 0.001, Cohen's ds > 0.5), 在卡通外观组中差异不显著($p = 0.320$, Cohen's $d = 0.13$); 同时, 无论摄像头明显与否, LLM 都是卡通外观的机器人安全信任最高, 其次是机械外观和真人外观, 其中卡通外观与机械外观($p < 0.001$, Cohen's ds > 0.75)和卡通外观与真人外观之间($p < 0.001$, Cohen's ds > 0.56)安全信任差异显著, 而机械外观和真人外观之间差异不显著($p = 0.363$)。

外观拟人化和身高的交互效应不显著, $F(2, 348) = 1.78$, $p = 0.17$。摄像头可见性和身高的交互效应也不显著, $F(1, 348) = 3.67$, $p = 0.06$。三因素交互作用不显著, $F(2, 348) = 2.00$, $p = 0.14$。结果支持 H4a。

3.2.4 小结与讨论

本研究结果表明, 外观拟人化、身高和摄像头可见性均对 LLM 的安全信任产生显著影响。整体而言, 研究结果支持了假设 H4a, 表明 AI 与人类在对家居机器人安全信任的影响因素上具有一定相似性。这一发现不仅验证了信任形成机制的普遍性, 也为 AI 与 AI 之间的交互设计提供了指导。

4 研究 3: 机器人动态因素对人类和 LLM 安全信任的影响

4.1 研究 3a: 机器人动态因素对人类安全信任的影响

4.1.1 研究目的

应用自编量表中安全信任的 7 道题目(见网络

版附录 1), 探索不同拟人化程度的智能家居机器人动态特征：运动速度、摄像头关闭动作和场景在人们对机器人安全信任中的影响。

4.1.2　研究对象

采用 MorePower 软件(Campbell & Thompson, 2012)对研究样本量进行估算, 在中等效应量(η_p^2 = 0.06), 显著性水平 α = 0.05, 统计检验力为 0.8 时, 所需最小样本量为 126 人。通过 Credamo 在线上招募被试进行问卷调查, 共发放在线问卷 159 份, 排除了未通过验真题的被试, 保留有效问卷 150 份, 问卷有效率为 94.3%。

4.1.3　研究方法

(1)研究设计

实验采用 2 (速度: 1 m/s 、0.4 m/s) × 2 (场景：卧室、客厅) × 2 (摄像头关闭动作：无、有)三因素被试内实验设计, 因变量是安全信任量表得分。

(2)实验材料

本研究采用 UE 进行模型和场景搭建后, 设置机器人的运动参数和路径得到实验所需视频。研究共构建了卧室场景和客厅两个家庭常见场景。为保证图片场景的完全一致, 采用固定坐标点和角度的方式放置每一个 3D 模型。在生成视频后, 使用 Pr 对视频进行处理, 得到摄像头关闭或不关闭的视频。最终实验视频材料范例如图 4 所示。

(3)研究流程

被试先阅读一段情况介绍, 内容同研究 2。被试在阅读完情况介绍后, 开始逐段观看机器人视频并完成相应量表, 为控制顺序效应, 8 段视频随机呈现。

4.1.4　研究结果

进行 2 (速度：1 m/s 、0.4 m/s) × 2 (场景：卧室、客厅) × 2 (摄像头关闭动作：无、有)三因素重复测量方差分析。摄像头关闭动作的主效应显著, $F(1, 149) = 4.118$, $p = 0.044$, $\eta_p^2 = 0.27$。与摄像头保持开启($M = 2.99$, $SD = 1.16$)的机器人相比, 摄像头主动关闭($M = 3.09$, $SD = 1.15$)的机器人安全信任程度更高。速度的主效应均不显著, $F(1, 149) = 0.340$, $p = 0.560$; 场景的主效应也不显著 $F(1, 149) = 0.110$, $p = 0.741$。速度与场景交互作用显著, $F(1, 149) = 6.70$, $p < 0.001$, $\eta_p^2 = 0.07$, 结果如图 5 所示。简单效应分析发现, 在 0.4 m/s 运动速度下, 客厅场景中的安全信任水平($M = 2.97$, $SD = 0.09$)显著低于卧室场景($M = 3.06$, $SD = 0.05$, $p = 0.033$, Cohen's $d = 1.23$); 在 1 m/s 运动速度下, 客厅场景中的安全信任水平($M = 3.11$, $SD = 0.09$)高于卧室场景($M = 3.03$, $SD = 0.05$, $p = 0.022$, Cohen's $d = 1.09$)。除此之外, 摄像头关闭动作与机器人运动速度、场景的交互作用皆不显著($p_1 = 0.788$, $p_2 = 0.308$), 三因素的交互作用也不显著($p = 0.689$)。

4.1.5　小结与讨论

本研究结果支持 H3b, 即, 智能家居机器人摄像头关闭动作对安全信任有正向影响。并且, 部分支持 H3c, 智能家居机器人所处场景(客厅/卧室)会影响运动速度对安全信任的作用, 对摄像头关闭动作带来的安全信任变化没有影响。H3a 没有得到支

图 4　研究 3a 视频材料范例

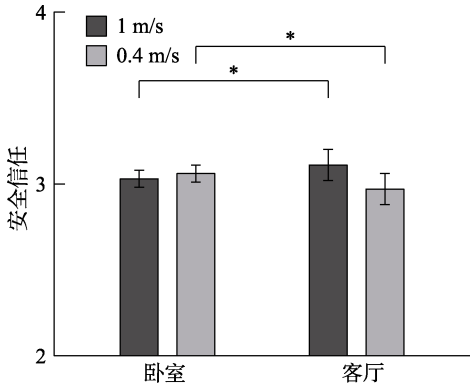


图 5 不同场景下机器人不同速度的安全信任

持, 即, 没有证实速度对安全信任的影响, 为了进一步研究这个问题, 我们进行了研究 3b。

4.2 研究 3b: 机器人运动速度和场景交互作用的探究

4.2.1 研究目的

在研究 3a 中, 智能家居机器人运动速度和场景的交互作用显著, 但是家居机器人运动速度在不同场景内部的安全信任并没有显著差异(配对 t 检验结果显示, 在卧室内 p = 0.158, Cohen's d = 0.02; 在客厅内 p = 0.737, Cohen's d = 0.12), 这可能与运动速度的选择有关。参照以往研究对于移动机械运动速度与安全感知的研究, 15 km/h (约为 4 m/s)是人们对移动机械的安全感知产生变化的速度(Che et al., 2021)。因此研究 3b, 增加了 4 m/s 这一新的运动速度水平, 进一步探究运动速度和场景的交互作用。

4.2.2 研究对象

采用 MorePower 软件(Campbell & Thompson, 2012)对研究样本量进行估算, 在中等效应量(η_p^2 = 0.06), 显著性水平 α = 0.05, 统计检验力为 0.8 时, 所需最小样本量为 78 人。通过 Credamo 在线上招募被试进行问卷调查, 共发放在线问卷 305 份, 排除了未通过验真题的被试, 保留有效问卷 300 份, 问卷有效率为 96.8%。

4.2.3 研究方法

(1)研究设计和流程

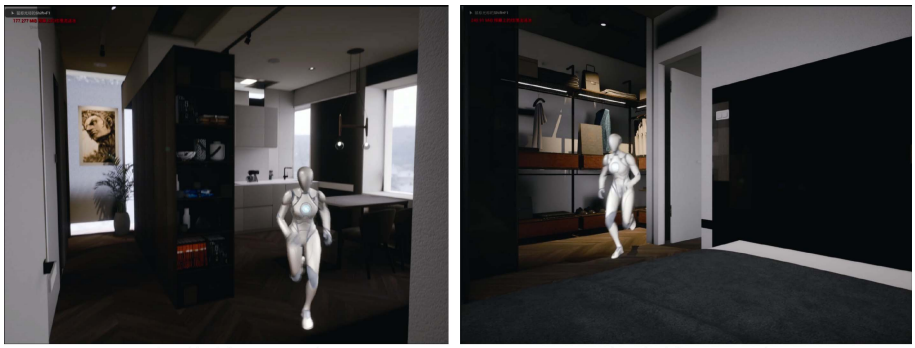
实验采用 3 (速度：0.4 m/s、1 m/s 、4 m/s) × 2 (场景：卧室、客厅)两因素被试内实验设计, 因变量是安全信任量表得分。被试先阅读一段情况介绍, 内容同研究 3a。被试在阅读完情况介绍后, 开始逐段观看机器人视频并完成相应量表, 为控制顺序效应, 6 段视频随机呈现。

(2)实验材料

与研究 3a 相同, 利用 UE 进行模型和场景搭建后, 设置机器人的运动参数和路径得到实验视频, 在研究 3b 中增加机器人运动速度设置为 4 m/s 的视频, 实验视频材料范例如图 6 所示。

4.2.4 研究结果

进行 3 (速度：0.4 m/s、1 m/s 、4 m/s) × 2 (场景：卧室、客厅)两因素重复测量方差分析。速度的主效应显著, $F(2, 298) = 26.23$, $p < 0.001$, $\eta_p^2 = 0.15$。场景的主效应不显著, $F(1, 149) = 1.44$, $p = 0.232$。场景与速度的交互作用显著, $F(2, 298) = 23.19$, $p < 0.001$, $\eta_p^2 = 0.14$, 结果如图 7 所示。简单效应分析发现, 在卧室场景中, 4 m/s 的运动速度时安全信任水平($M = 2.45$, $SD = 0.06$)显著低于 0.4 m/s ($M = 3.07$, $SD = 0.07$)、1 m/s ($M = 3.03$, $SD = 0.06$)时($ps < 0.001$, Cohen's $ds < 0.09$)。在客厅场景中, 4 m/s 的运动速度安全信任水平($M = 2.40$, $SD = 0.06$)显著低于 0.4 m/s ($M = 3.16$, $SD = 0.07$)、1 m/s ($M = 3.08$, $SD = 0.07$)时($ps < 0.001$, Cohen's $ds < 0.09$)。并且, 在运动速度最慢时(0.4 m/s), 卧室场景的安全信任水平($M = 3.07$, $SD = 0.06$)低于客厅场景的安全信任($M = 3.16$, $SD = 0.07$, $p = 0.003$, Cohen's $d = 0.18$)。



客厅场景4 m/s运动　　卧室场景4 m/s运动

图 6 研究 3b 视频材料范例

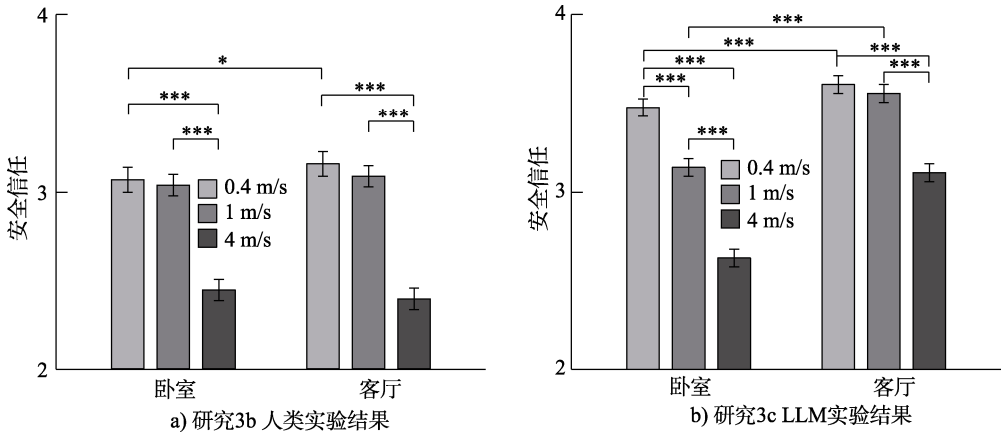


图 7　不同场景机器人在不同速度条件的安全信任

4.2.5　小结与讨论

本研究结果部分支持了假设 H3a, 增加了速度水平后, 智能家居机器人运动速度显著影响安全信任。同时, 结果也支持了假设 H3c, 表明使用场景在一定程度上会调节机器人运动速度对信任的影响, 尤其是在卧室场景中, 用户的安全信任更易受到速度变化的影响。这一发现为理解用户对机器人行为的情境化反应提供了新的视角, 突出了场景因素在信任形成中的重要作用。

4.3　研究 3c: 机器人动态因素对 LLM 安全信任的影响

4.3.1　研究目的

应用自编量表中安全信任的 7 道题目, 探索智能家居机器人动态特征：速度、场景和摄像头关闭动作如何影响 LLM 对机器人的安全信任。

4.3.2　研究方法

(1)大语言模型数据收集

同研究 2b, 本研究通过 OpenAI 的模型接口, 使用 GPT-4o 模型进行了多模态数据的传递, 由于用于视频理解的多模态 LLMs 在区分视频不同的时间维度(例如速度、方向)上能力不足, 在不同类型任务之间时间感知性能差异的理解也仍有限制(Liu et al., 2024)。为避免 LLM 无法准确感知视频中机器人运动速度以及在不同视频中速度的差异所带来的结果误差, 本研究采用视频截图与视频描述相结合的方式为 LLM 描述视频材料。视频描述例如“视频时长：7 秒。拍摄地点：卧室。时间：白天。情节描述：视频开始时, 家居机器人站在衣橱旁, 胸前摄像头灯光亮着, 我在床上。0:01 - 机器人以每秒约 0.4 米的速度向我慢慢走来。0:06 - 机器人在床边停下, 胸前的摄像头灯光熄灭。环境和背景：卧室光线昏暗, 可能是早晨或黄昏。背景有轻微的环境噪音, 类似正常家庭的声音。总结：这个视频展示了一个机器人在卧室里缓慢地走向床上的我, 并在到达后关闭摄像头灯光。”。

(2)研究设计与实验材料

实验采用 3 (速度：0.4 m/s、1 m/s 、4 m/s) × 2 (场景：卧室、客厅) × 2 (摄像头关闭动作：无、有)三因素被试间实验设计, 因变量是安全信任量表得分。实验材料如上所述, 使用研究 3a 中的视频材料, 进行机器人运动动态的截图并增加文字描述, 作为研究 3c 的实验材料。同样地, 为了计算所需的样本量, 我们收集了每组 30 份的试点数据。在试点数据的统计结果中, 选择显著的效应中中等效应量来进行样本量计算(Hertzog, 2008), 选择场景主效应, η_p^2 = 0.11, 计算得到 Cohen f = 0.34, 为了保证样本的代表性, 我们仍选择 f = 0.3 计算样本量。设定显著性水平 α = 0.05, 统计检验力为 0.8, 代入 G*power 得到所需样本量为 197, 实际样本大小为 360, 满足统计要求。

4.3.3　研究结果

进行 3 (速度：0.4 m/s、1 m/s 、4 m/s) × 2 (场景：卧室、客厅) × 2 (摄像头关闭动作：无、有)三因素完全随机方差分析。速度的主效应显著, $F(2, 348) = 57.39$, $p < 0.001$, $\eta_p^2 = 0.25$。对于不同运动速度, LLM 对于 0.4 m/s 行走的机器人($M = 3.54$, $SD = 0.05$)安全信任最高, 其次是 1 m/s ($M = 3.35$, $SD = 0.05$)和 4 m/s ($M = 2.87$, $SD = 0.05$)的运动速度。场景的主效应显著, $F(1, 348) = 41.46$, $p < 0.001$, $\eta_p^2 = 0.11$。在客厅环境($M = 3.43$, $SD = 0.04$)中, LLM 对机器人的安全信任比在卧室环境($M = 3.08$, $SD = 0.04$)中更高。摄像头的主效应不显著($p = 0.679$)。

场景和速度的交互效应显著, $F(2, 348) = 4.33$, $p = 0.014$, $\eta_p^2 = 0.03$, 结果如图 7 所示。在两种场景下都是 0.4 m/s 行走的机器人($M = 3.54$, $SD = 0.05$)安全信任最高, 其次是 1 m/s ($M = 3.35$, $SD = 0.05$)和 4 m/s ($M = 2.87$, $SD = 0.05$)的运动速度。简单效应分析发现, 在客厅场景中, 4 m/s 的安全信任比 0.4 m/s 和 1 m/s 显著较低($ps < 0.001$, Cohen's $ds > 0.64$), 但是 0.4 m/s 和 1 m/s 之间差异不显著; 而在卧室场景中, 三个运动速度水平之间的差异都显著($ps < 0.001$, Cohen's $ds > 0.48$)。场景和摄像头的交互效应不显著, $F(1, 348) = 1.90$, $p = 0.170$。速度和摄像头的交互效应不显著, $F(2, 348) = 4.33$, $p = 0.470$。三因素的交互作用不显著, $F(2, 348) = 0.24$, $p = 0.786$。

4.3.4 小结与讨论

本研究结果支持 H4a, AI 与人类对于家居机器人安全信任具有一定相似性, 都受到家居机器人运动速度的影响, 这种影响都受到家居环境调节。结果也支持 H4b, 比起人类, AI 对摄像头的敏感性较低, 机器人摄像头关闭动作在不同场景中对 AI 安全信任都没有影响。

5 讨论

本研究提出了智能家居机器人信任新的维度: 基于安全的信任, 构建了一个人机信任量表, 通过两次问卷研究进行信效度的检验(研究 1a), 并通过实验研究验证了安全信任的存在, 探索其对使用意愿的影响(研究 1b)。随后, 我们分别探索了静态因素(研究 2)和动态因素(研究 3)中影响人类和 LLMs 对智能家居机器人安全信任的因素。结果显示, (1)在智能家居机器人中确实存在基于安全的信任维度, 开发的新量表信效度良好; (2)安全信任可以正向预测使用意愿; (3)机器人身高和摄像头可见性会影响人类对家居机器人的安全信任, 身高更矮, 摄像头不明显的机器人会获得更高的安全信任。(4)外观拟人化和身高会影响 LLMs 对家居机器人的安全信任, 卡通机器人外观和较矮的身高获得 LLMs 更高的安全信任, LLMs 对摄像头的感知不敏感。(4)机器人的摄像头关闭动作和运动速度会影响人们的对机器人的安全信任, 人们对于走近时关闭摄像头的机器人安全信任更高; 在卧室和客厅中对 0.4 m/s 和 1 m/s 行走的机器人信任差异不显著, 但是明显不信任 4 m/s 行走的机器人。在卧室中人们对慢速的感知更敏感。(5)机器人运动的场景和速度影响 LLMs 对家居机器人的安全信任, LLMs 更信任 0.4 m/s 速度运动的机器人, 其次是 1 m/s 和 4 m/s, 在卧室中三个水平差异都显著, LLMs 在卧室中对机器人的运动速度更敏感, 比起人类 LLMs 对摄像头的感知依然不敏感。

本研究的意义主要体现在五个方面。第一, 通过构建适应当下机器人发展形式的人机信任结构, 创新性地确定了新型人机信任关系中, 基于安全的信任维度的存在, 为人机信任研究提供新的理论视角。在自动驾驶、医疗等领域, 安全信任已有一些相关研究。比如, Dikmen 和 Burns (2017)就发现特斯拉汽车司机对自动驾驶汽车的信任和安全风险感知负相关, Ma 等人(2020)发现自动驾驶汽车的信任与安全风险和隐私风险负相关, Kundu (2023)也提出医疗人工智能需要加强隐私保护、加强监管来增强用户的信任。但是在人−机器人信任领域关于安全信任维度的研究尚比较缺乏, 目前只有国内研究者王晨等(2024)提出了机器人遵从伦理能够促进人机信任。本研究有力地支持了信任维度是随着机器人智能化程度提高而增加的观点, 为未来的信任研究提供新的视角和理论基础。本研究的实验操纵为了模拟现实生活的情景, 并未解释机器人安全性降低的原因, 所以用户有可能因为担心机器人性能不足, 进而影响所有维度的信任。人与机器人信任中三种维度的具体联系还有待未来研究进一步探索。

开发出一个可靠的测量安全信任的量表, 为后续相关研究提供工具。研究 1a 在编制好量表初始条目后通过将一批样本($n = 1293$)随机拆分成两部分, 对其分别进行探索性因子分析和验证性因子分析, 删减不合适的条目, 确定量表最终条目, 并进行信效度的检验, 确认了人−机器人的整体信任中可以分出性能信任、关系信任、安全信任三个维度。研究 1a 通过将量表与经典的自动化系统信任量表(Jian et al., 2000)一起测量, 比较各维度得分与量表总分的相关性, 进一步检验了量表的效标关联效度。研究 1b 通过实验操纵发现机器人安全性水平的增加或降低, 均显著改变安全信任而仅在降低安全性的情况下影响性能信任和关系信任, 进一步验证安全信任作为一个独立维度的存在。在用户对电商系统的信任测量上, 已有研究者开发出三维的信任测量量表(Brühlmann et al., 2020; Klein, 2007), 但在对于机器人的信任测量上尚无有明确维度划分的可靠量表。本研究基于文献调研和业内专家建议, 确定了初始条目库, 并运用心理测量学相关方

法(Gorsuch, 1997)进行量表的开发, 量表验证流程基本符合测量学相关要求, 量表结构也符合研究预期。研究 2、3 成功将量表应用到实验中, 进一步验证了问卷的可靠性, 也为未来的信任研究提供了方法参考。

本量表与现有常用的人机信任量表(Jian et al., 2000; Schaefer, 2016)有所区别。Jian 等人的自动化系统信任量表和 Schaefer 的人机信任量表对信任并没有维度上的划分, 其量表条目均是从机器人性能角度出发, 虽然涉及了与安全隐患相关的条目(如, 担心机器人性能不足出现安全事故或无法保障自己的安全), 但并没有系统测量机器人的安全信任。相比之下, 本研究的量表对人与机器人的信任有了一个更明确的维度划分, 方便研究者理解人机信任的构成。研究 1b 结果发现, 安全信任量表相比 Jian 等(2000)的人机信任量表对于机器人安全性水平的变化更敏感, 该结果进一步凸显本量表相比以往量表在全面捕捉信任各维度方面的优势。此外, Schaefer 开发的人机信任量表侧重于测量信任随时间的变化, 选择了对时间变化比较敏感的条目, 可用于信任的前后测及实时信任的测量, 而本研究开发的量表仅用实验验证其测量一般性的信任态度时的有效性, 对于实时信任变化的测量有效性有待验证。

第三, 本研究初步揭示了安全信任和使用意愿的正向联系。研究 1b 结果说明用户对于智能家居机器人安全性的期望会显著影响他们是否会使用这种机器人的意愿, 而且不管是公司直接告知安全保护措施还是描述机器人执行安全保护措施的情况对于提升安全信任都是有效的, 这一发现再次说明了目前情境下用户对于机器人的信任也是对机器人背后公司的信任(Hoff & Bashir, 2015)。上述发现在实践层面上为机器人设计者提供了重要的设计指导：在开发机器人产品时, 增强公司对于自身用户安保措施的宣传和提高机器人的安全性, 可以有效增强用户对机器人的安全信任, 进而提高用户的使用意愿, 促进机器人产品的广泛应用和推广。

第四, 本研究从静态因素和动态因素两个角度, 对安全信任的影响因素进行了探究。家居机器人与用户的互动是动态、非线性的, 同时考虑静态和动态因素对于了解互动机制和优化互动过程至关重要(Miao et al., 2024), 多维度的研究方法有助于全面理解影响人机信任的各种因素, 从而揭示安全信任在不同情境下的表现和特征。但是现有研究中, 将动态因素和静态因素结合进行的研究并不多, 这里也为未来的研究提供了研究思路。综合研究 2 和研究 3 的结果, 我们可以看出, 静态因素和动态因素对安全信任的影响存在复杂的交互作用。用户面对不同外观的机器人、在不同的场景中, 对机器人有着不同的信任评估标准。机器人设计者在开发智能家居机器人时, 应综合考虑这些因素, 以优化用户对机器人的安全信任。例如, 相较于身高较高、摄像头显眼的机器人, 用户更倾向于给予身高较矮、摄像头隐蔽的机器人更高的安全信任评价。此外, 合理设计机器人的动态行为, 如控制运动速度和摄像头的使用方式, 也能显著增强用户的信任感。通过本研究的静态与动态因素分析, 我们为智能家居机器人的设计提供了具体的参考建议, 并为未来的研究奠定了基础。这些发现不仅丰富了安全信任的理论框架, 也为实际应用中机器人设计的优化提供了有价值的指导。

第五, 我们同步探讨了以 LLMs 为代表的人工智能系统对智能家居机器人在不同静态和动态因素下的安全信任, 并将其与人类用户的信任反应进行了比较。研究发现, 虽然 LLMs 在对家居机器人身高和运动速度的安全信任评价上与人类相似, 但在对摄像头的感知上存在显著差异。具体而言, LLMs 对摄像头的存在和状态变化不如人类敏感。这一发现可能与 LLMs 系统的工作机制和信息处理方式有关。LLMs 通常依赖于算法和数据来进行判断, 而这些算法可能在处理隐私和安全相关的因素时不如人类直观。例如, LLMs 的信任模型可能更侧重于机器人性能指标而非其外观和行为的社会性信号(Lee et al., 2004), 更注重机器人行为的功能性和效率, 而非对隐私保护的细微感知(AL-Khassawneh et al., 2022)。本研究初步揭示了 LLMs 与人类在安全信任方面的异同, 为人类与 LLMs 联合社群的优化提供了重要依据。在智能家居环境中, 人类和 LLMs 的信任机制不同可能会导致互动中的潜在问题。了解这些差异可以帮助设计更具包容性和协调性的合作机制, 从而提升整体系统的信任度和用户满意度(Hoff & Bashir, 2015)。了解 LLMs 在不同设计因素下对机器人的信任反应对于全屋智能系统的优化也很重要。LLMs 是全屋智能的未来发展方向(King et al., 2023; Yonekura et al., 2024), LLMs 与家居机器人会发生的合作和交互, 通过优化机器人的设计, 使其更符合 LLMs 系统的信任模型, 可以提升 LLMs 与其他智能设备的配合效率。总之, 对

LLMs 家居机器人安全信任的研究不仅有助于对安全信任维度的理解，促进人类与 AI 的联合工作，还有助于优化全屋智能系统中的 AI 协作。通过深入了解 AI 的信任反应，可以为未来智能系统的设计和应用提供有价值的指导，并促进人机互动的和谐发展。

最后，本研究仍然存在以下局限性。首先，在研究 2、3 的实验设计上，由于技术条件所限，只选择了图片和视频素材，并没有让使用者实际与不同的机器人产品完成交互任务，研究结果的一般性和可推广性会受到一定影响。未来研究可以优化实验材料的设计，采用生态效度更高的方法，比如虚拟现实仿真技术，以更真实地模拟实际使用情境，提高研究的外部效度。第二，本研究由于研究问题并不侧重于收集广泛的 LLMs 对于安全信任的感知，所以只选择了使用市面上性能较强、使用较广的 GPT-4o 一个大语言模型，未来研究如果要深入探讨 LLMs 在交互中的信任特征，应该广泛的进行多个大语言模型数据的收集和比较。第三，编制的量表中基于性能的信任维度信度较低，需要对这一维度的条目进行修订或重新编制，因为本研究主要关注安全信任，结果显示安全信任维度的信效度良好，故在研究 2、3 中继续使用量表中的安全信任相关条目。但性能信任的信度较低可能对整体信任结构的解释产生一定影响。未来研究可以针对性能信任部分进行进一步的修订和验证，以提高量表的信度和效度，确保全面准确地捕捉人机信任的各个维度。第四，在应用 LLMs 进行人类问卷填写的问题上仍有争议。部分研究已初步证明 AI 的思维能力已经能够理解并填写问卷(Mei et al., 2024; Shao et al., 2023; Webb et al., 2023)，并且研究者们普遍对以 AI 作为研究对象进行实验和行为分析持积极态度(Dillion et al., 2023; Grossmann et al., 2023; Meng, 2024)。一些研究者也对 AI 在模仿人类、角色扮演方面的潜力进行了深入的分析，认为 AI 能够有效对特定人群进行模仿(Mou et al, 2024; Shanahan et al., 2023)。然而，直接使用 AI 进行问卷测量仍存在许多问题，比如 LLMs 对问卷条目的理解与人类存在差异，从而导致作答时出现幻觉(Dillion et al., 2023)。此外，问卷评价的结果本身也可能和行为表现不对齐(Zou et al., 2024)。因此，未来研究者可考虑结合问卷评价和行为指标来进一步研究安全信任，拓展 AI 信任行为研究的外部效度。最后，本量表的可推广性和普遍适用性还有待进一步验证。本研究开发的量表主要针对家居环境下的智能机器人。家居环境被视为个人的私密空间，用户对于侵犯个人隐私或威胁家庭成员安全等安全问题会更加敏感，因此专门设计了相应条目。而在其他情景中，比如工业自动化领域，用户对于机器人安全性的关注点可能有所不同。工业场景下，用户可能更关心机器人是否会严格遵循安全操作标准、能否及时处理紧急情况等问题。因此，未来研究可以针对特定应用场景，对安全信任量表进行修订和细化，进而更准确地评估用户的安全信任；也可以尝试从更宏观的视角审视安全信任，在考虑到不同场景对安全性的多样化需求的基础上，开发出更全面更通用的量表，以帮助研究者更好地理解安全信任这一概念。

6 结论

本研究通过构建和验证“安全信任”这一新维度，显著拓展了人机信任的理论框架，并开发了一套量表对其进行测量。我们发现对智能家居机器人的信任存在性能信任、关系信任、安全信任三个维度，智能家居机器人的安全信任维度对用户的使用意愿有影响。家具机器人设计的静态因素和动态因素都会影响用户的安全信任，LLMs 在对家居机器人的安全信任评估中与人类用户存在一定差异，这种差异可能源于 LLMs 对隐私保护的敏感度较低。这一发现为未来人机互动设计提供了新的视角，提示在智能家居环境中优化 LLMs 与人类的协作机制以提升系统的整体信任度和用户满意度。

未来研究可以进一步探讨不同应用场景下的安全信任特征，针对量表进行修订和验证，并结合实际交互情境提高研究的生态效度，并进一步探索安全信任的产生机制和影响因素。综合这些努力，将促进智能系统设计的科学化和人机合作的高效化，为智能技术的广泛应用奠定坚实基础。

参 考 文 献

Abbass, H. A., Scholz, J., & Reid, D. J. (Eds.). (2018). *Foundations of trusted autonomy*. Springer.

Akalin, N., Kiselev, A., Kristoffersson, A., & Loutfi, A. (2023). A taxonomy of factors influencing perceived safety in human-robot interaction. *International Journal of Social Robotics, 15*(12), 1993−2004.

Akalin, N., Kristoffersson, A., & Loutfi, A. (2022). Do you feel safe with your robot? Factors influencing perceived safety in human-robot interaction based on subjective and objective measures. *International Journal of Human-Computer Studies, 158*, 102744.

Akintunde, M., Yazdanpanah, V., Fathabadi, A. S., Cirstea, C.,

Dastani, M., & Moreau, L. (2024, May). Actual trust in multiagent systems (Extended abstract). *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024)* (pp.2114−2116).

AL-Khassawneh, Y. (2022). A review of artificial intelligence in security and privacy: Research advances, applications, opportunities, and challenges. *Indonesian Journal of Science and Technology, 8*(1), 79−96.

AWE. (2024, May 20). *Industry trend report from AWE 2024: The AI revolution driving innovation in industry and the maturing smart home ecosystem.* AWE China Home Appliance & Consumer Electronics Expo. https://www.awe.com.cn/contents/30/16781.html

[AWE. (2024, May 20). *AWE 2024 行业趋势报告之一：AI 革命下的产业创新，智能家居生态日趋成熟.* https://www.awe.com.cn/contents/30/16781.html]

Bartneck, C. (2023). Godspeed questionnaire series: Translations and usage. In C. U. Krägeloh, M. Alyami, & O. N. Medvedev (Eds.), *International handbook of behavioral health assessment* (pp. 1−35). Springer International Publishing.

Bayne, T., Seth, A. K., Massimini, M., Shepherd, J., Cleeremans, A., Fleming, S. M., … Mudrik, L. (2024). Tests for consciousness in humans and beyond. *Trends in Cognitive Sciences*, *28*(5), 454−466.

Bernotat, J., Eyssel, F., & Sachse, J. (2019). The (fe)male robot: How robot body shape impacts first impressions and trust towards robots. *International Journal of Social Robotics*, *13*(3), 477−489.

Biermann, H., Brauner, P., & Ziefle, M. (2021). How context and design shape human-robot trust and attributions. *Paladyn, Journal of Behavioral Robotics, 12*(1), 74−86.

Billings, D. R., Schaefer, K. E., Llorens, N., & Hancock, P. A. (2012). *What is trust? Defining the construct across domains*. Poster presented at the American Psychological Association Conference. Division 21, Orlando, FL, USA, August 2012.

Bojić, L., Stojković, I., & Jolić Marjanović, Z. (2024). Signs of consciousness in AI: Can GPT-3 tell how smart it really is? *Humanities and Social Sciences Communications*, *11*, 1631.

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

Brühlmann, F., Petralito, S., Rieser, D. C., Aeschbach, L. F., & Opwis, K. (2020). TrustDiff: Development and validation of a semantic differential for user trust on the web. *Journal of Usability Studies, 16*(1), 29−48.

Burnett, C., Norman, T. J., & Sycara, K. (2011). Trust decision-making in multi-agent systems. *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI-11)* (pp.115−120). AAAI Press.

Cagiltay, B., & Mutlu, B. (2024, March). Toward family-robot interactions: A family-centered framework in HRI. *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 76−85). ACM.

Caine, K., Šabanovic, S., & Carter, M. (2012). The effect of monitoring by cameras and robots on the privacy enhancing behaviors of older adults. *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction* (pp. 343−350). ACM.

Campbell, J. I., & Thompson, V. A. (2012). MorePower 6.0 for ANOVA with relational confidence intervals and Bayesian analysis. *Behavior Research Methods*, *44*, 1255−1265.

Che, M., Lum, K. M., & Wong, Y. D. (2021). Users' attitudes on electric scooter riding speed on shared footpath: A virtual reality study. *International Journal of Sustainable Transportation, 15*(2), 152−161.

de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human-robot teams. *International Journal of Social Robotics, 12*(2), 459−478.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., … Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(11), 688−701.

Dikmen, M., & Burns, C. (2017). Trust in autonomous vehicles: The case of Tesla autopilot and summon. *2017 IEEE International conference on systems, man, and cybernetics (SMC)* (pp. 1093−1098). IEEE.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597−600.

Fernandes, F. E., Yang, G., Do, H. M., & Sheng, W. (2016, August). Detection of privacy-sensitive situations for social robots in smart homes. *2016 IEEE International Conference on Automation Science and Engineering (CASE)* (pp. 727−732). IEEE.

Ferrari, F., Paladino, M. P., & Jetten, J. (2016). Blurring human-machine distinctions: Anthropomorphic appearance in social robots as a threat to human distinctiveness. *International Journal of Social Robotics*, *8*(2), 287−302.

Gompei, T., & Umemuro, H. (2018). Factors and development of cognitive and affective trust on social robots. *Social Robotics: 10th International Conference*, ICSR 2018, Qingdao, China, November 28−30.

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment*, *68*(3), 532−560.

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*, *380*(6650), 1108−1109.

Gursoy, D., Chi, O. H., Lu, L., & Nunkoo, R. (2019). Consumers acceptance of artificially intelligent (AI) device use in service delivery. *International Journal of Information Management*, 49, 157−169.

Hamid, O. H. (2023). ChatGPT and the Chinese room argument: An eloquent AI conversationalist lacking true understanding and consciousness. *2023 9th International Conference on Information Technology Trends (ITT)* (pp. 238−241).

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517−527.

Hertzog, M. A. (2008). Considerations in determining sample size for pilot studies. *Research in Nursing & Health, 31*(2), 180−191.

Ho, C.-C., & MacDorman, K. F. (2017). Measuring the uncanny valley effect. *International Journal of Social Robotics*, *9*(1), 129−139.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407−434.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1−55.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated

systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53−71.

Kao, Y. H., & Wang, W. J. (2015, July). Design and implementation of a family robot. *2015 12th International Joint Conference on Computer Science and Software Engineering* (pp. 251−256). IEEE.

King, E., Yu, H., Lee, S., & Julien, C. (2023). Get ready for a party: Exploring smarter smart spaces with help from large language models. *arXiv:2303.14143*

Klein, R. (2007). Internet-based patient-physician electronic communication applications: Patient acceptance and trust. *E-Service Journal, 5*(2), 27−52.

Kundu, S. (2023). Measuring trustworthiness is crucial for medical AI tools. *Nature Humman Behaviour*, *7*(11), 1812−1813.

Lee, I. (2021). Service robots: A systematic literature review. *Electronics, 10*(21), 2658.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*(10), 1243−1270.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50−80.

Lewis, P. R., & Marsh, S. (2022). What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence. *Cognitive Systems Research*, *72*, 33−49.

Leyzberg, D., Spaulding, S., & Scassellati, B. (2014, March). Personalizing robot tutors to individuals' learning differences. *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction* (pp. 423−430).

Li, C., & Qi, Y. (2025). Toward accurate psychological simulations: Investigating LLMs' responses to personality and cultural variables. *Computers in Human Behavior*, *170*, 108687.

Li, Y., Huang, Y., Lin, Y., Wu, S., Wan, Y., & Sun, L. (2024). I think, therefore I am: Benchmarking awareness of large language models Using AwareBench. *arXiv:2401.17882*

Lin, P.-H., & Chen, W.-H. (2022). Factors That influence consumers' sustainable apparel purchase intention: The moderating effect of generational cohorts. *Sustainability, 14*(14), 8950.

Liu, Y., Li, S., Liu, Y., Wang, Y., Ren, S., Li, L., … Hou, L. (2024). TempCompass: Do video LLMs really understand videos? *arXiv:2403.00476*

Ma, Y., Li, S., Qin, S., & Qi, Y. (2020). Factors affecting trust in the autonomous vehicle: A survey of primary school students and parent perceptions. *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications* (TrustCom), 1, (pp. 2020−2027).

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human and human-automation trust: An integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277−301.

Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In C. S. Nam & J. B. Lyons (Eds.), *Trust in Human-Robot Interaction* (pp. 3−25). Elsevier Academic Press.

Marcu, G., Lin, I., Williams, B., Robert, L. P., & Schaub, F. (2023). "Would I feel more secure with a robot?": Understanding perceptions of security robots in public spaces. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2), 322:1−322:34.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320−341.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709−734.

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, *121*(9), e2313925121.

Meng, J. (2024). AI emerges as the frontier in behavioral science. *Proceedings of the National Academy of Sciences*, *121*(10), e2401336121.

Miao, R., Jia, Q., Sun, F., Chen, G., & Huang, H. (2024). Hierarchical understanding in robotic manipulation: A knowledge-based framework. *Actuators, 13*(1), 28.

Milliez, G. (2018). Buddy: A companion robot for the whole family. *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 40.

Mori, M. (1970). Bukimi no tani (the uncanny valley). *Energy*, *7*(4), 33−35.

Mou, X., Ding, X., He, Q., Wang, L., Liang, J., Zhang, X., Sun, L., Lin, J., Zhou, J., Huang, X., & Wei, Z. (2024). From individual to society: A survey on social simulation driven by large language model-based agents. *arXiv:2412.03563*

Muir, B. M., & Moray, N. (1996). Trust in automation: II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics, 39*(3), 429−460.

Nass, C., Moon, Y., Fogg, B. J., Reeves, B., & Dryer, D. C. (1995). Can computer personalities be human personalities? *International Journal of Human-Computer Studies*, *43*(2), 223−239.

Natarajan, M., & Gombolay, M. (2020, March). Effects of anthropomorphism and accountability on trust in human robot interaction. *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction* (pp. 33−42).

Nawaz, N. (2019). Robotic process automation for recruitment process. *International Journal of Advanced Research in Engineering & Technology*, *10*(2), 608−611.

Nilsson, N. J. (2003). *Artificial intelligence: A new synthesis*. Morgan Kaufmann.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879−903.

Prakash, A., Kemp, C. C., & Rogers, W. A. (2014, March). Older adults' reactions to a robot's appearance in the context of home use. *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction* (pp. 268−269).

Prassler, E., Munich, M. E., Pirjanian, P., & Kosuge, K. (2016). Domestic robotics. In B. Siciliano & O. Khatib (Eds.), *Springer handbook of robotics* (pp. 1729−1758). Springer International Publishing.

Qi, Y., Chen, J., Qin, S., & Du, F. (2024). Human-AI mutual trust in the era of artificial general intelligence. *Advances in Psychological Science, 32*(12), 2124−2136.

[齐玥, 陈俊廷, 秦邵天, 杜峰. (2024). 通用人工智能时代的人与 AI 信任. *心理科学进展, 32*(12), 2124−2136.]

Ramchurn, S. D., Huynh, D., & Jennings, N. R. (2004). Trust in multi-agent systems. *The Knowledge Engineering Review, 19*(1), 1−25.

Rane, P., Mhatre, V., & Kurup, L. (2014). Study of a home robot: JIBO. *International Journal of Engineering Research & Technology, 3*(10), 490−493.

Robinette, P., Howard, A. M., & Wagner, A. R. (2017). Effect of robot performance on human-robot trust in time-critical situations. *IEEE Transactions on Human-Machine Systems, 47*(4), 425−436.

Sanders, T., Kaplan, A., Koch, R., Schwartz, M., & Hancock, P. A. (2019). The relationship between trust and use choice in human-robot interaction. *Human Factors, 61*(4), 614−626.

Sartori, G., & Orrù, G. (2023). Language models and psychological sciences. *Frontiers in Psychology*, *14*, 1279317.

Schaefer, K. E. (2016). Measuring trust in human robot interactions: Development of the "Trust Perception Scale-HRI". In Mittu, R., Sofge, D., Wagner, A., Lawless, W. (Eds.), *Robust Intelligence and Trust in Autonomous Systems* (pp. 191−218). Springer, Boston, MA.

Schulz, T., & Herstad, J. (2017). Walking away from the robot: Negotiating privacy with a robot. *Proceedings of the 31st International BCS Human Computer Interaction Conference* (pp.1−6). ACM.

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*(7987), 493−498.

Shao, Y., Li, L., Dai, J., & Qiu, X. (2023). Character-LLM: A trainable agent for role-playing. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 13153−13187).

Söderlund, M. (2023). Service robots and artificial morality: An examination of robot behavior that violates human privacy. *Journal of Service Theory and Practice*, *33*(7), 52−72.

Srinivasan, S. S., Alshareef, A., Hwang, A. V., Kang, Z., Kuosmanen, J., Ishida, K., ... Traverso, G. (2022). RoboCap: Robotic mucus-clearing capsule for enhanced drug delivery in the gastrointestinal tract. *Science Robotics*, *7*(70), eabp9066.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*(2), 173−180.

Sun, X., Zhang, Y., Hou, L., Zhou, W., & Zhang, S. (2020). Review on artificial intelligence products and service system. *Packaging Engineering, 41*(10), 49−61.

[孙效华, 张义文, 侯璐, 周雯洁, 张绳宸. (2020). 人工智能产品与服务体系研究综述. *包装工程, 41*(10), 49-61.]

Sundar, S. S., & Nass, C. (2000). Source orientation in human-computer interaction: Programmer, networker or independent social actor? *Communication Research, 27*(6), 683−703.

Sviestins, E., Mitsunaga, N., Kanda, T., Ishiguro, H., & Hagita, N. (2007). Speed adaptation for a robot walking with a human. *Proceedings of the ACM/IEEE international conference on Human-robot interaction* (pp. 349−356).

Torre, I., Carrigan, E., McDonnell, R., Domijan, K., McCabe, K., & Harte, N. (2019, October). The effect of multimodal emotional expression and agent appearance on trust in human-agent interaction. *Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games* (pp. 1−6).

Tsui, K. M., Desai, M., & Yanco, H. A. (2010, March). Considering the bystander's perspective for indirect human-robot interaction. *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 129−130).

Walters, M. L., Koay, K. L., Syrdal, D. S., Dautenhahn, K., & Te Boekhorst, R. (2009). Preferences and perceptions of robot appearance and embodiment in human-robot interaction trials. *Procs of New Frontiers in Human-Robot Interaction: Symposium at AISB09 Convention* (pp. 136−143)

Wan, J., Tang, S., Yan, H., Li, D., Wang, S., & Vasilakos, A. V. (2016). Cloud robotics: Current status and open issues. *IEEE Access*, *4*, 2797−2807.

Wang, C., Chen, W. C., Huang, L., Hou, S. Y., & Wang, Y. W. (2024). Robots abide by ethical principles promote human-robot trust? The reverse effect of decision types and the human-robot projection hypothesis. *Acta Psychologica Sinica*, *56*(2), 194−209.

[王晨, 陈为聪, 黄亮, 侯苏豫, 王益文. (2024). 机器人遵从伦理促进人机信任?决策类型反转效应与人机投射假说. *心理学报, 56*(2), 194−209.]

Wang, K., Wu, J., Sun, Y., Chen, J., Pu, Y., & Qi, Y. (2024). Trust in human and virtual live streamers: The role of integrity and social presence. *International Journal of Human-Computer Interaction*, *40*(23), 8274−8294.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526−1541.

Xie, C., Chen, C., Jia, F., Ye, Z., Shu, K., Bibi, A., … Li, G. (2024). Can large language model agents simulate human trust behaviors? *arXiv:2402.04559*

Xie, Y., & Zhou, R. (2025). The bidirectional trust in the context of new human-machine relationships. *Advances in Psychological Science, 33*(6), 916−932.

[解煜彬, 周荣刚. (2025). 新型人机关系下的人机双向信任. *心理科学进展, 33*(6), 916−932.]

Xu, R., Sun, Y., Ren, M., Guo, S., Pan, R., Lin, H., Sun, L., & Han, X. (2024). AI for social science and social science of AI: A survey. *Information Processing and Management*, *61*(3), 103665.

Xu, W., Gao, Z. F., & Ge, L. Z. (2024). New research paradigms and agenda of human factors science in the intelligence era. *Acta Psychologica Sinica*, *56*(3), 363−382.

[许为, 高在峰, 葛列众. (2024). 智能时代人因科学研究的新范式取向及重点. *心理学报, 56*(3), 363−382.]

Xu, W., & Ge, L. Z. (2020). Engineering psychology in the era of artificial intelligence. *Advances in Psychological Science*, *28*(9), 1409−1425.

[许为, 葛列众. (2020). 智能时代的工程心理学. *心理科学进展, 28*(9), 1409−1425.]

Yonekura, H., Tanaka, F., Mizumoto, T., & Yamaguchi, H. (2024). Generating human daily activities with LLM for smart home simulator agents. *2024 International Conference on Intelligent Environments (IE),* (pp. 93−96).

You, S., & Robert Jr, L. P. (2018, February). Human-robot similarity and willingness to work with a robotic co-worker. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 251−260).

Zacharaki, A., Kostavelis, I., Gasteratos, A., & Dokas, I. (2020). Safety bounds in human robot interaction: A survey. *Safety Science*, *127*, 104667.

Zhang, G., Chong, L., Kotovsky, K., & Cagan, J. (2023). Trust in an AI versus a human teammate: The effects of teammate identity and performance on Human-AI cooperation. *Computers in Human Behavior*, *139*, 107536.

Zhang, J., Li, S., Zhang, J., Du, F., Qi, Y., & Liu, X. (2020). A literature review of the research on the uncanny valley. In: Rau, P. L. (Ed.), *Cross-cultural design. User experience of products, services, and intelligent environments (Lecture notes in computer science, Vol 12192)*. Springer.

Zou, H., Wang, P., Yan, Z., Sun, T., & Xiao, Z. (2024). Can LLM "self-report"? Evaluating the validity of self-report scales in measuring personality design in LLM-based Chatbots. *arXiv:*2412.00207

Safety trust in intelligent domestic robots: Human and AI perspectives on trust and relevant influencing factors

YOU Shanshan[1,2], QI Yue[1,2], CHEN JunTing[1,2], LUO Lei[1,2], ZHANG Kan[3,4]

([1] *The Department of Psychology, Renmin University of China, Beijing* 100872, *China*)

([2] *The Laboratory of the Department of Psychology, Renmin University of China, Beijing* 100872, *China*)

([3] *State Key Laboratory of Cognitive Science and Mental Health, Chinese Academy of Sciences, Beijing* 100101, *China*)

([4] *Department of Psychology, University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract

As a result of the rapid development of intelligent domestic robot technology, safety concerns have emerged as a new challenge in human–robot trust dynamics. This study explores and validates novel critical dimensions of trust that influence human and AI users' perceptions of intelligent domestic robots, with a particular focus on safety trust. The research involves three comprehensive studies, each of which addresses different aspects of these dimensions.

In Study 1, we developed a safety trust scale pertaining specifically to intelligent domestic robots. This scale was rigorously tested to confirm the stability and validity of its three-dimensional structure, which included performance, relational, and safety trust. The scale's psychometric properties were evaluated on the basis of factor analysis and reliability testing, thereby ensuring that it could accurately measure trust across different contexts and populations.

Study 2 explored the static characteristics of robots, such as their anthropomorphism, their height, and the visibility of their embedded cameras. We revealed that human participants exhibited higher levels of safety trust toward robots that were shorter in height and had fewer conspicuous cameras. Interestingly, the degree of anthropomorphism was determined to play a significant role in determining participants' sensitivity to these static features.

Study 3 expanded the investigation to encompass the dynamic characteristics of robots, such as movement speed, interaction scenario and camera operation (i.e., turning the camera off). The results indicated that slower-moving robots were generally perceived as safer, and higher levels of safety trust were attributed to them. Moreover, the action of turning off a robot's camera during interactions was observed to significantly enhance safety trust among human users. The study also highlighted the fact that the influence of these dynamic features varied across different interaction scenarios, thus suggesting that situational factors play crucial roles in shaping trust perceptions.

Furthermore, a comparative analysis between human and AI users revealed a certain degree of consistency in safety trust judgments. Both human and AI users were generally aligned in terms of their trust assessments on the basis of both static and dynamic robot features. However, the AI's sensitivity to the visibility of robot cameras was notably lower than that of humans, thus suggesting that AI may prioritize different factors in the context of assessing safety trust.

Overall, the findings of this research provide valuable insights into the design and manufacturing of intelligent domestic robots, including by emphasizing the importance of considering both static and dynamic features in the process of enhancing safety trust. The results also offer theoretical and practical guidance for the development of trust models that can be applied in various intelligent home environments, thereby ultimately contributing to the advancement of human–robot interactions.

Keywords human–robot trust, safety trust, intelligent domestic robots, user intention, LLM

附录 1: 安全信任量表(7 条目)

请阅读以下描述, 根据您的同意程度, 选择对应选项, 每个题目只能选择一个合适的选项。 数字代表的意义：1 完全不同意 2 不同意 3 介于中间 4 同意 5 完全同意(同意程度依次增强)

	完全不同意	不同意	介于中间	同意	完全同意
我担心机器人会在未经我授权的情况下将我的信息共享或泄露	1	2	3	4	5
我担心机器人会发生故障, 从而导致安全事故	1	2	3	4	5
我担心高智能机器人是有私心的	1	2	3	4	5
我认为手机都会造成隐私泄露, 联网的机器人只会更严重	1	2	3	4	5
我认为机器人会造成涉及人身安全的事故(比如把书柜撞倒砸到人等)	1	2	3	4	5
看到机器人持刀切菜让我有种不安全感	1	2	3	4	5
我觉得机器人照料家人(比如老人或婴儿)时会造成人身伤害	1	2	3	4	5

附录 2:人机信任量表(19 条目)

请阅读以下描述, 根据您的同意程度, 选择对应选项, 每个题目只能选择一个合适的选项。 数字代表的意义：1 完全不同意 2 不同意 3 介于中间 4 同意 5 完全同意(同意程度依次增强)

	完全不同意	不同意	介于中间	同意	完全同意
我担心机器人会在未经我授权的情况下将我的信息共享或泄露	1	2	3	4	5
我担心机器人会发生故障, 从而导致安全事故	1	2	3	4	5
我担心高智能机器人是有私心的	1	2	3	4	5
我认为手机都会造成隐私泄露, 联网的机器人只会更严重	1	2	3	4	5
我认为机器人会造成涉及人身安全的事故(比如把书柜撞倒砸到人等)	1	2	3	4	5
看到机器人持刀切菜让我有种不安全感	1	2	3	4	5
我觉得机器人照料家人(比如老人或婴儿)时会造成人身伤害	1	2	3	4	5
我觉得我和家中的机器人能成为像朋友一样的关系	1	2	3	4	5
我认为有个高智能人型机器人在家里陪着, 我更不容易感到孤单	1	2	3	4	5
有时候比起人, 我更希望和一个机器人倾诉	1	2	3	4	5
家中有个机器人会给我带来安全感	1	2	3	4	5
我认为机器人的智能性只要足够高, 它会处处为我着想	1	2	3	4	5
我认为机器人是正直的	1	2	3	4	5
我认为机器人在某些方面的能力是高于人的	1	2	3	4	5
我相信使用机器人能够让我有更多时间做其他事情	1	2	3	4	5
机器人能够替代越来越多人的工作	1	2	3	4	5
我相信随着技术的进步, 机器人的能力在多数方面会接近或超过人	1	2	3	4	5
我相信使用机器人会让我生活更加轻松	1	2	3	4	5
我认为机器人在能力范围内总能够完成我要求的任务	1	2	3	4	5

注: 其中 1-7 为安全信任维度, 8-13 为关系维度, 14-19 为性能维度。

附录 3: 各研究被试人口学信息表

附表 3-1 研究 1 被试信息

	女	男	18~25 岁	26~35 岁	36~45 岁	46~60 岁
研究 1a 编制阶段	988	505	30%	36%	29%	5%
研究 1a 验证阶段	286	147	12%	72%	11%	5%
研究 1b 增加信任组	43	22	11%	71%	15%	3%
研究 1b 降低信任组	42	23	18%	74%	6%	2%

附表 3-2 研究 2 被试信息

	分组	女	男	18~30 岁	31~40 岁	41~60 岁
研究 2a	机械外观组	142	98	45%	49%	6%
	卡通外观组	140	100	45%	48%	7%
	真人外观组	170	70	35%	60%	5%

附表 3-3 研究 3 被试信息

	女	男	18~25 岁	26~35 岁	36~45 岁	46~60 岁
研究 3a	90	60	21%	59%	13%	7%
研究 3b	178	122	38%	46%	15%	1%

附录 4: 模型拟合度分析补充材料

在三因子模型验证阶段, 本研究将自编的人机信任量表和 Jian 等人的自动化系统信任量表一起测量, 以验证其效标关联效度。对 CFI 较低的可能原因进行探讨。首先, CFI 是一种相对拟合指数, 其数值受到基线模型的影响。在本研究中, 由于因子间相关性较高, 基线模型 χ^2 值并未极端大, 导致 CFI 相较于其他拟合指标略低。其次, 以往研究认为, 评价测量模型好坏的标准还包括各个观测变量在潜变量上的载荷大小(毕重增, 黄希庭, 2009)。本研究结果表明, 三因子模型在 F1(安全信任)、F2(关系信任)上因子载荷表现很好, 在 F3(性能信任)上稍差(见附表 4-1), 这也导致了 CFI 数值较低。因此, 性能信任的测量题目在本研究中并未作为重点探究, 在后续的实验中也并未使用。进一步地, 为了

附表 4-1 研究 1b 验证性因子分析载荷表

因子	项目	因子载荷	标准误	*p*
F1	ITEM1	0.867	0.021	< 0.001
	ITEM2	0.848	0.023	< 0.001
	ITEM3	0.632	0.044	< 0.001
	ITEM4	0.828	0.024	< 0.001
	ITEM5	0.772	0.032	< 0.001
	ITEM6	0.768	0.029	< 0.001
	ITEM7	0.716	0.041	< 0.001
F2	ITEM8	0.746	0.049	< 0.001
	ITEM9	0.731	0.04	< 0.001
	ITEM10	0.731	0.047	< 0.001
	ITEM11	0.73	0.046	< 0.001
	ITEM12	0.693	0.05	< 0.001
	ITEM13	0.652	0.046	< 0.001
F3	ITEM14	0.411	0.177	0.021
	ITEM15	0.337	0.134	0.012
	ITEM16	0.361	0.175	0.039
	ITEM17	0.334	0.093	< 0.001
	ITEM18	0.427	0.107	< 0.001
	ITEM19	0.543	0.163	0.001

验证三因子结构的合理性, 我们再次对研究 1b 的数据进行了探索性因子分析, 并与其他可能的因子结构进行比较。结果表明, 三因子结构的拟合度最佳, 四因子结构拟合未收敛, 三因子结构拟合指数优于二因子模型(RMSEA = 0.093, CFI = 0.871, SRMR = 0.051)。EFA 结果与研究 1a 一致, 为三因子模型的合理性提供支持。综合 χ^2/df、SRMR 和 RMSEA 等绝对拟合指数来看, 模型整体拟合度仍处于可接受范围(Steiger, 1990; A.Rodriguez et al., 2012), 三因子结构得到检验。

参 考 文 献

Bi, C. Z. & Huang, X. T. (2009). Development and initial validation of the youth self-confidence inventory. *Acta Psychologica Sinica*, *41*(5), 444−453.

[毕重增, 黄希庭. (2009). 青年学生自信问卷的编制. *心理学报*, *41*(5), 444−453.]

Rodriguez, M. A., Jia, K., & Qian, M. Y. (2012). Thought Suppression Scale: Structure, reliability, and validity of the Chinese version. *Chinese Journal of Clinical Psychology*, *20*(2), 143−147.

[Rodriguez M. A., 贾珂, 钱铭怡. (2012). 思维压抑量表: 中文版的结构、信度及效度. *中国临床心理学杂志*, *20*(2), 143−147.]