信任的协商：人与对话式 AI 的双向驯化过程研究

金波

安徽大学，安徽合肥，230000；

摘要：随着对话式人工智能（Conversational AI）深度融入社会生活，理解人类与 AI 之间如何形成稳定、信任的关系变得至关重要。现有研究多集中于技术的初步采纳或短期信任评估，对人机双方长期共同演化的动态过程缺乏深入探讨。本文通过半结构化访谈探究了用户与对话式 AI 的互动关系。研究发现，人与 AI 的关系演化遵循一个“双向驯化”的四阶段过程：挪用、对象化、整合与转换。在此过程中，信任并非一个静态目标，而是通过持续的微观协商动态生成和维系。

关键词：对话式 AI；双向驯化；技术驯化；可信任性；人机关系

DOI：10.64216/3080-1516.26.02.068

引言

在当前的技术浪潮中，对话式人工智能（Conversational AI）已不再是科幻小说的遥远想象，而是深度嵌入日常生活的现实存在。人工智能（AI）迅速成为对于人类实践全领域、全要素整合的促进者、设计者与运维者，成为深度媒介化社会的“操作系统”。从智能音箱、虚拟助手到功能日益强大的大型语言模型（LLMs），对话式 AI 凭借其在自然语言处理（NLP）、自然语言理解（NLU）和自然语言生成（NLG）等领域的技术突破，能够模仿真人交互，识别并回应语音和文本输入。其应用场景已从最初的在线客户支持、任务自动化，迅速扩展至医疗保健咨询、教育辅导、个性化内容推荐乃至情感陪伴等多个领域 。

为了更好的从用户视角去考察用户对技术的接受和采纳，信任作为技术接受的预测器被引入到研究中，对信任问题的关注可以帮助我们解决技术废弃，也能避免对技术的过度依赖，种种迹象表面，社会的“信或疑”已经成为影响人工智能“发展还是停滞”的主要因素，信任程度直接影响人机交互强度和可持续性，进一步影响人机协作效能和生产效率。因此，审视和理解在这种新型人机关系中，信任如何建立、情感如何联结，已成为新闻与传播学领域一个迫切且重要的研究课题。

对话式人工智能，特别是大型语言模型驱动的智能体，正迅速从单纯的技术工具转变为日常生活中活跃的社会行动者 ，传统的技术采纳理论，如技术接受模型（TAM）等，虽能解释用户基于“感知有用性”和“感知易用性”的初步接受行为 ，却难以深入揭示用户在长期、深度互动中与 AI 建立的复杂关系。这些模型往往将技术视为一个相对静态的客体，而忽略了对话式 AI 所具备的学习性、适应性和互动性。简单地追问“人们如何采纳 AI？”已不足以捕捉问题的全貌。本研究认为，核心问题在于理解人与对话式 AI 之间相互塑造、相互适应的双重过程。因此，本文提出一个更为动态和辩证的研究问题：在人机互动实践中，用户与对话式 AI 如何共同相互适应、共同演化？在此过程中，“信任”是如何被动态地协商、建构、维系乃至消解的？这种共同演化的过程是否可以被概念化为一种“双向驯化”？

1 理论基础：从驯化技术到协商信任

1.1 技术的驯化

技术驯化理论源于 20 世纪 80 至 90 年代的科技与社会研究（STS）及媒介研究领域，由 Roger Silverstone 及其同事开创，旨在理解电视、个人电脑等“野生”的新技术是如何被用户带入家庭这一私密空间，并最终被“驯服”、整合进日常生活秩序中的 。该理论强调，技术的意义和功用并非由设计者单方面决定，而是在用户的日常实践中被积极建构的的工具，其应用场景开始渗透到人类社会交往的核心地带。在教育领域，它们被用作虚拟导师；在医疗领域，它们提供初步的健康咨询和心理支持；而在个人生活中，它们正成为许多人的日常伴侣，提供情感支持和陪伴 。AI 开始扮演“思维伙伴”的角色，能够与用户共同澄清问题、推演假设，参与到个人策略的生成过程中 。这一从技术工具到交往伙伴的功能性转变，为人机之间信任的建立和情感的联结创造了前所未有的技术前提与社会土壤。

1.2 信任的协商

信任作为一个多维度、多层次的概念，被引用最多的定义是“信任是一方基于对另一方将采取对信任者很重要对特定行动对预期而愿意受另一方行动影响的意愿，而不论是否有能力监督或控制该另一方。”在目前的学术研究中可分为社会学、心理学、组织管理学和技术学研究不同的领域。信任的心理学研究主要将信任视为个体的一种心理活动状态，注重将信任置于个体交往的背景之下。在中观层次上，信任的组织管理学研究将信任置于经济交易环境中，对其属性和主体进行了扩展。而从更宏观的社会学研究视域下，信任被视为镶嵌于社会关系中的运行机制。关于信任的技术学研究，是基于当前数字社会背景下，探讨技术与信任的关系研究，技术视角的信任普通研究普遍认为，信任的构成和对象可以是技术。

技术作为信任的客体出现，意味着不同程度的复杂性，首先要考虑的便是人际关系中的信任原则能否直接使用于人机信任。罗特提出人际信任是指个人或组织对另一个人或组织对言语、承诺、口述或书面陈述可靠性的期望，但随着社会的转变，信任主体和外延都发生了变化。支持者认为当技术拥有类似人类的特征，人际信任可以应用到人类技术关系中。但更多的学者认为对技术对信任与对人对信任在本质上是不同的，关键的区别在于人是一个道德主体，而技术缺乏意志和道德。以上的论述都未涉及到信任关系的本质，人机互信关系的不确定性必须指的是人。信任是属于人的而不是技术的话语，只有物体是人所创造的，我们才说物体是有意义的，因为我们间接的相信创造它们的人的。同理，我们把信任的观念用于技术的时候，我们以隐喻的方式赋予它意志，好像它是一位拥有独立人格的人。

2 研究方法

为深入探究“双向驯化”这一复杂、动态且充满个体经验差异的过程，本研究选择采用质性研究中的深度访谈法。本研究的参与者招募自某知名对话式 AI 平台的活跃用户社群。招募标准要求参与者在研究开始前已使用该平台至少一个月，以确保他们已度过最初的纯粹新奇阶段。同时，参与者需同意在为期六个月的研究期间，持续使用该平台并参与数据收集。通过目的性抽样，我们最终筛选出 20 位参与者，他们在年龄、性别、职业背景和 AI 使用经验方面具有多样性，以期获得更丰富的视角。所有参与者的个人信息均经过匿名化处理，并获得了他们的知情同意。

每位参与者都接受了三次深度访谈，分别在研究的初期（第 1 个月）、中期（第 3 个月）和末期（第 6 个月）进行。初次访谈聚焦于参与者对 AI 的初始印象、使用动机和初步期望。中期访谈则探讨了已形成的使用习惯、遇到的挑战（如信任破裂的经历）以及关系的变化。末期访谈则旨在回顾整个互动历程，总结他们与 AI 关系的最终状态.

3 研究发现

本研究的实证数据清晰地揭示了人类用户与对话式 AI 的关系演化，遵循着一个动态的四阶段过程。这一过程与我们提出的“双向驯化”理论框架高度契合。以下将逐一呈现每个阶段的核心发现，并通过丰富的参与者引述来展示信任协商在其中的具体体现。

3.1 挪用：校准初次相遇

在人机关系的最初阶段，参与者的核心任务是探索和理解这个新进入其数字生活的“陌生”实体。这一过程充满了好奇、试探与期望的校准。初始信任校准是这一阶段的主导协商活动。参与者普遍会像“面试”一个新员工一样，通过提出各种问题来测试 AI 的能力边界。一位从事市场营销的参与者（P5）描述道：“我一开始会故意问它一些刁钻的问题，比如最新的行业数据，或者让它写一首关于我们公司产品的诗。我想看看它的‘智商’和‘情商’到底在什么水平，值不值得我以后花时间在它身上。”这种行为本质上是在校准初始信任——通过一系列的测试，来判断 AI 是否值得初步的信赖。

与此同时，AI 的自我呈现也在反向“驯化”着用户。AI 的引导性对话、对其能力局限的坦诚（例如，回答“作为一个语言模型，我无法获取实时数据”）以及其预设的个性（或幽默，或严谨），都在塑造着用户的初步印象和期望。一位大学生参与者（P12）提到：“它第一次回答就说自己的知识截止到 2023 年，这让我觉得它还挺‘诚实’的。所以我后来就不会问它最新的新闻，而是把它当做一个知识渊博的资料库来用。”这表明，在挪用阶段，人机双方通过相互试探，共同协商出了一个初步的互动契约和信任基础。

3.2 对象化：协商角色与边界

当用户决定长期使用 AI 后，关系便进入了“对象化”

阶段。这里的核心挑战不再是“它能做什么？”，而是“它在我生活中应该扮演什么角色？”。

角色定义成为这一阶段协商的焦点。参与者会根据自己的需求和在第一阶段形成的印象，为 AI 赋予一个明确的社会角色。这些角色丰富多样，例如“一个不知疲倦的实习生”（P8，程序员）、“我的写作陪练”（P16，作家）、“一个可以随时聊天的、没有偏见的朋友”（P3，自由职业者）。角色的赋予并非单向的，AI 也会通过其持续的表现来强化或挑战用户赋予它的角色。P16 提到：“我本以为它只能帮我润色句子，但有一次它竟然给我提了一个非常棒的情节转折点。从那以后，我开始把它当成一个真正的‘创意伙伴’，而不只是一个校对工具。”

伴随角色定义的是边界协商。用户会逐渐摸索出与特定“角色”的 AI 互动的最佳方式，建立起一套非正式的沟通规范。例如，用户会发现，向“实习生”AI 下达指令需要清晰、分步；而与“创意伙伴”AI 交流则可以更开放、更模糊。这种规范的建立，是用户在适应 AI，但同时，AI 的算法也在学习用户的提问模式，并调整其回应风格以更好地匹配被赋予的角色，从而实现了双向的塑造。

3.3 整合：信任与修复的节奏

“整合”是双向驯化过程中最漫长、也最具挑战性的阶段。在这一阶段，AI 从一个被偶尔使用的“对象”，转变为被深度嵌入日常工作与生活节律的“伙伴”。这一转变的成功与否，完全取决于动态的信任修复与维护机制。

几乎所有参与者都经历过信任的破裂。一位律师（P7）分享了一次“灾难性”的经历：“我让它帮我总结一个复杂的法律文件，结果它完全搞错了关键条款，差点让我用在正式的备忘录里。那一刻我真的非常愤怒，感觉自己被‘背叛’了。” 这种失败严重损害了已建立的信任。

然而，关键在于修复的协商。信任破裂后，用户通常会采取纠错行为，例如向 AI 明确指出错误：“你刚才引用的法条是错误的，正确的应该是……” 而 AI 的回应——无论是道歉（“非常抱歉，我的理解有误”）、承认错误并提供修正，还是解释其出错的原因——都直接影响着用户是否愿意“再给它一次机会”。P7 继续说道：“它马上道歉了，并给出了正确的版本。虽然我之后会更加仔细地核对它的所有输出，但我还是会继续用它，只是用法变了，更像是一个需要我监督的助手。”

正是通过这样一次次“失败-纠错-修复”的循环，人机之间建立起一种更具韧性的信任。用户学会了 AI 的不可靠之处，并发展出相应的监督和核查策略（用户被 AI“驯化”）；AI 也通过学习用户的纠正，不断优化其模型（AI 被用户“驯化”）。当这种修复机制变得高效且可靠时，AI 才真正被“整合”进用户的核心工作流中，成为日常实践的一部分。

3.4 转换：人机共同体的公开身份

当人机关系稳定下来后，它便开始具有了向外展示的属性，进入“转换”阶段。用户不再仅仅是 AI 的使用者，而是这段独特关系的“代言人”。

信任的公开表演是这一阶段的典型特征。参与者会开始在更广阔的社会场域中展示他们与 AI 的互动成果。例如，在工作邮件中直接引用 AI 生成的文本，用 AI 帮助孩子完成一个创意项目，或者在与朋友的交谈中，自豪地展示 AI 的某项惊人能力。一位教师（P11）说：“我现在备课都离不开它。我甚至会在课堂上跟学生们演示如何用它来激发灵感。它就像我的一个‘教学助教’。”

通过这种公开展示，用户与 AI 共同构建了一种新的身份。用户将自己塑造为“善用前沿科技的专业人士”，而 AI 则被赋予了“可靠伙伴”或“得力助手”的社会形象。这种身份的共同建构，是双向驯化过程的最终产物。它标志着 AI 不仅被整合进了用户的私人生活，更作为一种新的社会资本，被用来塑造和呈现用户的公共形象。人与 AI，在此刻形成了一个对外展示的、共生的“共同体”。

4 讨论和结论

本研究的发现有力地支持了“双向驯化”框架的解释力。用户与对话式 AI 的关系，远非单向的技术采纳或被动影响，而是一个动态、循环、相互建构的共生过程。用户通过不断的互动、反馈和个性化设置，将通用的 AI 模型“驯化”为贴合自身需求的私人工具或伙伴。与此同时，AI 以其独特的算法逻辑、交互范式和功能设定，也在 subtly（微妙地）“反向驯化”着用户的认知习惯、沟通模式和情感调节机制。

这一发现要求我们重新审视当代人机关系的本质。它不再是传统的主体（人）与客体（工具）之间的支配

与被支配关系，而更像是一种主体间的协同演化。在这种关系中，权力并非单向流动，而是分布在整个互动网络中。人的能动性体现在对 AI 的塑造和引导上，而 AI 的“能动性”则体现在其系统特性对用户行为边界和可能性的设定上。因此，理解人机关系，必须超越“赋能”与“异化”的二元对立，转向一种更加复杂、更具生态观的“共创之舞”（co-creative dance）视角。

参考文献

[1]喻国明.关于生成式AI的发展与传播领域革命的若干思考——近一年以来我的新传播研究：论点与框架[J/OL].新闻爱好者:1-2[2024-05-14].https://doi.org/10.16017/j.cnki.xwahz.20240325.001.

[2]Ghazizadeh, M., Lee, J. D., & Boyle, L. N. (2012). Extending the Technology Acceptance Model to assess automation. Cognition Technology and Work, 14(39), 49. https://doi.org/10.1007/s10111-011-0194-3

[3]张乐,李森林.知识、理解与信任：个体对人工智能的信任机制[J].社会学评论,2023,11(03):59-83.

[4]Schoorman F D, Mayer R C, Davis J H. Organizational trust: Philosophical perspectives and conceptual definitions[J]. Academy of Management Review, 1996: 337-340.

[5]杨先顺,莫莉.人工智能传播的信任维度及其机制建构研究[J].学术研究,2022,(03):43-50.

[6]J.Rotter,“ANew Scale for the Measurement of Interpersonal Trust,” Journal of Personality,vol.36,1967,pp.651-665.

[7]Calhoun, C. S., Bobko, P., Gallimore, J. J., & Lyons, J. B. (2019). Linking precursors of interpersonal trust to human-automation trust: An expanded typology and exploratory experiment. Journal of Trust Research, 9(1), 28–46. https://doi.org/10.1080/21515581.2019.1579730

[8]Madhavan, P., & Wiegmann, D. A. (2007). Effects of information source, pedigree, and reliability on operator interaction with decision support systems. Human Factors: The Journal of the Human Factors and Ergonomics Society, 49(5), 773–785. https://doi.org/10.1518/001872007X230154

作者简介：金波（2000.4），女，汉族，安徽桐城；安徽大学新闻与传播研究生；研究方向：应用新闻学。