心理科学进展 2024, Vol. 32, No. 12, 2124–2136
Advances in Psychological Science

https://doi.org/10.3724/SP.J.1042.2024.02124

通用人工智能时代的人与 AI 信任*

齐 玥[1,2] 陈俊廷[1,2] 秦邵天[1,2] 杜 峰[3,4]

([1]中国人民大学心理学系, 北京 100872) ([2]中国人民大学心理学系实验室, 北京 100872)
([3]中国科学院行为科学重点实验室, 北京 100101) ([4]中国科学院大学心理学系, 北京 100049)

摘 要 随着技术的发展, 通用人工智能初见雏形, 人机交互以及人机关系将进入新的时代。人与人工智能(AI)的信任关系也即将从单方向的人对 AI 信任逐渐转变为人与 AI 的互信。本研究在回顾社会心理学中的人际信任模型与工程心理学中的人机信任模型的基础上, 从人际信任视角提出了人与 AI 动态互信模型。该模型将人与 AI 视为对等的信任建立方, 结合信任与被信任方的影响因素、结果反馈和行为调整构建了人与 AI 动态互信的基本理论框架, 强调了人与 AI 信任中关系维度的“互信”与时程维度的“动态”这两个重要特征。模型首次将 AI 对人的信任以及二者互信的动态交互过程纳入分析, 为人与 AI 的信任研究提供新的理论视角。未来研究应更多关注 AI 对人的信任如何建立与维持、人与 AI 互信的量化模型以及多智能体交互中的人与 AI 互信。

关键词 信任, 人机互信, 信任校准, 人机关系, 人与 AI
分类号 B849

随着新一代信息技术的快速发展, 人工智能(AI)技术已经渗透到我们日常工作、生活的多个领域(de Visser et al., 2016), 从手机的智能助手, 到路上的无人驾驶汽车, 再到新一代聊天机器人, 人工智能已经不再是冷冰冰的机器系统, 而是成为人们日常生活、学习和工作中的助手(Walter et al., 2014)、同伴(Glikson & Woolley, 2020), 甚至恋人(Sullins, 2010), 扮演着愈加重要的角色。2023 年以来, 随着 ChatGPT 走入公众视野, 研究者们发现, AI 已经变得越来越像人(Binz & Eric, 2023; Shiffrin & Mitchell, 2023), 并提出最新的人工智能模型 GPT-4 是迈向通用人工智能(AGI)的重要一步(Bubeck et al., 2023), 人与 AI 的关系即将从工具的使用关系转变为协作关系。在人机协作中, AI 技术的成熟水平是前提, 但人类是否信任 AI 则成为调节人与 AI 之间协作的关键因素(Glikson & Woolley, 2020; 许为, 葛列众, 2020)。

信任是人与人工智能交互中的核心(Frison et al., 2019; Wright et al., 2003), 直接影响交互的成败以及用户的使用感受。比如, 在自动驾驶系统评估中信任影响着用户体验(Frison et al., 2019; Rödel et al., 2014)。究其原因, 一方面, 随着技术的发展, AI 算法变得愈加复杂, 形成“黑箱”。人们可以判断输入黑箱的数据和输出的结果, 却不知道里面发生了什么(Frison et al., 2019; Wright et al., 2003)。这使得用户难以理解其决策过程(Siau & Wang, 2020; Wang & Siau, 2019), 也难以预测人工智能的最终决策。此时, 用户对人工智能的信任将决定用户是否会使用人工智能算法的结果。另一方面, 为了提高人工智能的性能, 用户必须向人工智能系统提供个人数据(Stephanidis et al., 2019), 这种情况可能会导致用户隐私泄露的相关风险。因此, 用户是否信任人工智能并愿意将个人数据委托给人工智能系统是影响用户使用意愿的一个关键前提。

维持适当的信任水平也会影响人与 AI 的互

收稿日期: 2024-01-29
* 国家自然科学基金(32471130; 32000771; 32371107), 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)。
通信作者: 齐玥, E-mail: qiy@ruc.edu.cn
杜峰, E-mail: duf@psych.ac.cn

动结果, 即协作任务的完成质量。以自动驾驶为例, 信任是影响自动驾驶中人机协同效率与驾驶安全的关键要素(Hancock et al., 2019; Rahwan et al., 2019)。若驾驶员不信任 AI 系统, 可能会忽视自动驾驶系统提供的辅助功能, 无法有效降低疲劳驾驶、分心等风险驾驶行为; 相反, 若驾驶员过度信任 AI 系统, 则会完全放弃对行驶车辆的监控, 忽视自动驾驶系统的局限性, 从而导致巨大的交通安全隐患(Noah et al., 2016; Noah et al., 2017; Wintersberger et al., 2018)。在军事领域, 人与 AI 队友之间的信任, 对于团队任务完成至关重要(Groom & Nass, 2007)。无人机等人工智能系统在人机协同作战中的广泛应用, 也使得研究者越来越重视人与 AI 的信任关系(Chen et al., 2011)。伴随着通用人工智能时代的到来, 人机信任已成为人与人工智能是否和谐共处、协作发展的基础。

当前, 人与 AI 的交互关系已经开始转变, 但是现有的人与 AI 信任研究并没有准确理解这种新型的信任关系。这种理解不足主要体现在三个方面: 首先, 现有研究对于人与 AI 信任的定义并不明确, 这会导致不同研究者对人与 AI 信任的理解和应用存在差异; 其次, 传统的信任模型大多从人际信任和人机信任两个角度分别展开阐述, 但随着人工智能技术的提高, 人与 AI 的交互将逐步贴近人与人的交互, 融合心理学中两个不同的研究领域变得更加有价值; 最后, 现有的信任模型仅关注到人对 AI 的信任, 忽视了 AI 对人的信任这一角度, 对人机互动的双向信任过程缺乏理解。为解决现有研究的局限性, 本文将围绕人与 AI 信任的定义、信任模型的发展展开, 提出并阐述人与 AI 动态互信模型, 并在最后对人与 AI 信任的未来研究进行了展望。

为获得人与 AI 信任的相关文献, 在准备本综述的过程中, 本文采用的文献检索策略如下。本文在中国知网、Web of Science、IEEE Xplore、Elsevier、ScienceDirect 中进行关键词检索, 所使用的检索关键词包括“人机信任(Human-Machine Trust)”、“人工智能信任(trust in AI 或 trust in artificial intelligence)”、“自动化信任(trust in automation)”和“机器人信任(trust in robot)”, 文献检索的时间为 1994 年至 2024 年 1 月, 文献类型包括期刊论文和会议论文, 以确保涵盖近 30 年的研究成果。

1　人与 AI 信任的定义

信任是很多学科领域共同的研究主题, 在心理学、社会学、哲学、政治学、经济学等领域得到广泛研究。信任是一个复杂而模糊的概念, 不同的研究领域已给出超过 300 个信任定义。不一致的信任描述会导致研究者们无法在先前研究的基础上建立人与 AI 信任的研究体系, 因此, 对人与 AI 信任的明确定义对于研究人与 AI 的信任具有重要的理论和实践意义。本文将在综述过往相关领域研究和定义的基础上提出人与 AI 信任的定义。

在人机信任领域, Lee 和 See (2004)提出的定义被广为接受(Hoff & Bashir, 2015; Khastgir et al., 2017)。他们从态度角度定义信任, 提出脆弱性和不确定性是信任的前提, 并将人机信任定义为“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”。随后, 在人与自动化系统的交互中, 研究者开始提出自动化信任。例如, Billings 等(2012)回顾了 282 个信任定义, 包括 200 个人际信任定义以及 50 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的期望、信心、风险、脆弱性、依赖、态度及合作等特征。这些信任定义揭示了在人与自动化的合作关系中完成某项任务时所需的自动化信任的三项核心特征。首先, 在信任主体上, 必须包含信任关系的双方, 即有一个委托者(操作人员/用户)来给予信任, 有一个受托者(自动化)来接受信任; 其次, 双方所要共同完成的事情存在一定的风险, 必须存在受托人无法执行并完成任务的可能性, 从而引发不确定性和风险(Hardin, 2002); 最后, 受托者(自动化)必须具有执行并完成任务的动机及能力。

自动化信任是人与 AI 信任的前身, 在以往有关人机信任乃至人机交互的研究中, 自动化和 AI 常常混淆使用(Glikson & Woolley, 2020)。自动化是指计算机遵循预先编程的规则, 执行以前由人类执行的重复和单调任务的情况(Parasuraman & Riley, 1997)。传统自动化产生的行为及其结果是预先编程的, 因此, 用户能很好理解自动化系统的决策。传统的自动化是确定性的, 不包括任何学习过程(Raj & Seamans, 2019)。而人工智能不仅可以实现自动化, 比如, 机器学习算法可以制定

自动化过程遵循的规则，它们还可以根据经验和反馈进行学习和调整。总结上述有关自动化信任的要素，可以看出，自动化信任是建立在不确定的合作关系中，作为委托者和受托者相互协作完成任务的必要条件存在的。这种对于自动化信任的要素构建同时适用于人与人的信任关系，以及人与 AI 的信任关系。

但与自动化信任不同，人与 AI 信任关系的建立可能在无意中发生。在很多人工智能系统的使用场景中，人们甚至没有意识到 AI 的存在。比如，在嵌入式 AI 的研究中发现，人们可能没有意识到他们正在使用一个人工智能支持算法的应用程序。在一项针对 Facebook 用户进行的调查中研究者发现，超过一半(62%)的用户不知道有一种人工智能算法正在管理页面上的信息，决定将哪些信息呈现给他们，哪些信息应该隐藏(Eslami et al., 2015)。尽管参加研究的 Facebook 用户由于没有被告知 AI 算法的使用而感到不愉快、惊讶甚至愤怒，但在了解算法的工作原理后，用户仍在继续使用该平台。隐藏 AI 算法的使用并没有对用户的长期信任产生重大影响。可见，人对 AI 的信任并不一定受到使用前是否知情的影响。

更重要的是，随着 AI 智能化水平提高，人与智能系统的关系，将从单向度的人对 AI 信任，逐步转化为双向度的人与 AI 互信(许为，葛列众，2020)。因此，人与 AI 信任本身的定义也应与时俱进。伴随信任定义的更新也将衍生出一系列新的研究问题。基于此，本文提出人与 AI 信任的新定义，即无论是否意识到 AI 算法的存在，人们与 AI 系统之间所持有的认为对方能帮助自己实现特定目标的态度和信心，以及在互动过程中接受对方的不确定和脆弱性并为之承担相应风险的意愿。

本文的新定义综合了以往人机信任和自动化信任定义的内容，不仅涵盖 Lee 和 See (2004)所提出的基于态度的人机信任观点，也符合 Billings 等人(2012)总结的自动化信任三项核心特征：两个信任主体、完成的事情存在风险以及受托人有完成任务的动机和能力。在综合以往观点的基础上，新定义充分考虑了当今人与 AI 互动的特点：一方面针对 AI 技术使用的隐蔽性强调定义可以扩展到用户未意识到 AI 参与的情况，另一方面考虑到人与 AI 信任角色的转变，提出人与 AI 存在互信的关系，即信任包括用户作为委托者对 AI 的信任，也包括了 AI 作为委托者对用户输入的依赖和适应。这种互信关系也潜在揭示了人与 AI 信任的动态过程，交互过程中人与 AI 都会作为委托者，并根据受托者的行为来不断校准自己对受托者的信任。

2 人与 AI 信任的模型发展：从人−人到人−AI

人与 AI 信任的研究起源于人际信任。随着科技发展，人们越来越多地面临着与 AI 的互动，比如医疗领域(Forcier et al., 2020)、社交领域(Bartneck & Forlizzi, 2004)。在社交平台的聊天窗口中，人们已经难以仅仅从交互设计、互动形式和内容上区分出对面是人类还是 AI。在已经习惯于与人类合作的领域中，人们开始越来越多地将原先的互动对象从人类替换到 AI，由此产生了一个问题：我们是否像信任人类一样信任 AI？因此，在已知信任已成为人与 AI 有效互动的基本前提下(Hancock et al., 2011; van Pinxteren et al., 2019)，借鉴人际信任的理论来研究人与 AI 的信任是十分有价值的。目前，已有很多研究将人际互动的相关理论和模型转化为人机交互(HCI)和人−机器人交互(HRI)研究的理论和模型(例如，Aly & Tapus, 2016; de Visser et al., 2016; Gockley et al., 2006; Kulms & Kopp, 2018)。研究者曾将人类的刻板印象模型(Stereotype Content Model, Fiske et al., 1999, 2002)迁移到人机研究中。结果发现，能力与温暖可以正向影响人们对机器人的信任(Christoforakos et al., 2021)。因此，这些人际信任的决定因素可以迁移到人−机器人的交互以及人与 AI 的信任发展中。

2.1 人际信任模型

在人际信任领域，研究者提出，信任，本质上就是将自己认为重要的事情交由他人手中的一种选择，是对我们无法控制的风险的应对(Cofta, 2007; Deutsch, 1962)。在 Mayer 等人(1995)提出的信任模型中，人际信任的判断考虑了三个主要特征：潜在受托人完成委托者需要他们做的事情的能力，他们在决定是否做这件事时的仁慈，以及他们在尊重委托者并且就他们是否会做这件事达成的任何协议方面的正直。

早期 Mayer 等人的信任模型更多关心的是受托者一方的关键特征，随着对信任研究的深入，

McKnight 和 Chervany (1996)在 Mayer 等人(1995)的基础上, 结合信任决策过程, 提出了信任概念关系模型。他们指出, 尽管受托人的风险和可信度相当, 人们在某些情况下往往比在其他情况下更信任他人。因此, 新的模型强调, 信任的倾向不仅来自于受托人的特征, 也来自于潜在委托人的态度(如乐观主义)以及决策情境。这里的决策情境是指做出信任决策的更广泛的(即非个人的)社会情境, 即个体的信任决策还受到了系统信任的影响。例如, 一些社会文化比其他社会文化更倾向于培养人与人之间的普遍信任, 某些制度和社会规范的存在可能会导致系统信任倾向的增加或减少(Luhmann, 1990)。与 Mayer 早期的信任模型仅仅强调受托人的特质不同, 信任概念关系模型对于委托人和情境的强调为人机信任三维模型的提出描绘了合适的理论框架。

2.2　人机信任的四因素模型

2011 年, Sanders 等人针对人与机器人的信任提出了一个四因素模型, 将机器人性能、机器人依赖性、个体差异和协作作为人与机器人交互(Human-Robot Interaction, HRI)过程中的信任影响因素(Sanders et al., 2011)。这些影响因素, 从机器人属性(如拟人性、动物性、亲和力、感知智能和感知安全)到人类互动因素(如可用性、社会接受度、用户体验和社会影响), 广泛总结了以往研究中信任的前因变量。在该工作的基础上, 基于元分析结果, 研究者首次将影响人与机器人信任的前因变量总结为三个因素: 机器人相关因素(包括机器人绩效以及特性), 人类相关因素(包括能力和个人特征), 以及环境相关因素(包括团队协作和任务相关因素) (Hancock et al., 2011)。这项工作为之后经典的三因素模型的提出奠定了基础。

2.3　人机信任的三因素模型

2014 年, Schaefer 等人在人−机器人信任模型(Hancock et al., 2011)的基础上, 通过回顾人机信任相关文献, 发展出了人机信任的三因素模型(Schaefer et al., 2014)。其将对人机信任的影响因素分为操作者因素、机器系统因素和环境因素三类, 并进一步将与操作者相关的因素分为操作者特质、操作者状态、认知因素和情感因素这四种类型; 将与机器系统相关的因素进一步分类为机器系统特性和机器系统能力; 将与环境相关的因素分类为与任务相关和与团队相关。

在此基础上, 我国研究者提出了影响人工智能信任的三因素模型。人对 AI 的信任与以下三个方面相关: 与操作者个体的特征相关, 因为对人工智能的信任发端于人; 与情境特征相关, 良好的信任社会体系与信任社会制度将为人工智能的信任提供良好的存在语境; 与人工智能系统特征相关, 如技术的性能与效应等技术要素, 是人工智能信任的必备要素(闫宏秀, 2019; 高在峰 等, 2021)。简而言之, 人对人工智能信任的影响因素应包括个体、技术与环境。

2.4　人对 AI 信任的整合模型

2022 年, Lewis 和 Marsh 在针对 AI 信任研究的综述中提出了一个整合模型。该模型不仅适用于同伴之间的信任, 而且适用于启发式信任决策, 这一模型为人与 AI 动态互信模型的建立奠定了基础。

该模型提出, 人们的可信度判断高度依赖于可用信息的数量和种类。可用信息影响着人们对 4 个主要可信度特征的感知, 包括能力、可预测性、诚实与正直、意愿与仁慈(Lewis & Marsh, 2022)。受托人具备完成相应任务的能力, 行为一致能够被预测, 愿意履行承诺, 具有满足需求的意愿, 则信任决策能得到回应。当以上 4 个主要特征的信息难以得到时, 人们还能够通过代理信任(即对其他相关方的信任)来完成决策。比如, 对新产品的信任, 可能受到代理——生产厂商的影响。生产厂商的信誉度和其过往产品的质量会影响消费者对于新产品的可信度感知。在实际的人与 AI 关系中, 一个 AI 系统可能是具备能力的, 但是人们难以揣测它是否正直、仁慈, 因此, 实际上, 人们在大多数情况下这些影响因素会交织在一起, 共同影响人们对于 AI 的信任。

然而, 整合模型更多关注受托人, 即 AI 自身的特征对于感知可信度以及信任决策、行为的影响, 而忽略了用户状态的影响。

2.5　过往信任模型的综合比较与分析

通过上述模型可以看到研究者对于信任的理解是在不断演进和深化的。早期的信任模型是在人际互动情景下讨论的, Mayer 等人(1995)的模型开创性提出信任取决于受托人的能力、仁慈和正直三个特征, 但是却仅考虑到受托人一方的特征。McKnight 和 Chervany (1996)的模型拓展了委托人和情境两方面因素的影响, 更全面地解释了

人际信任的影响因素，也为人机信任模型的提出提供了大体框架。在人机交互领域，Sanders 等(2011)最早对人机信任研究进行总结，提出了人机信任的四因素模型，但是该模型并不能广泛概括所有影响信任的前因变量。Hancock 等人(2011)对现有模型进行了修订，将前因变量总结为三类因素，该模型仅基于人与机器人交互的相关研究，该领域里研究人类相关因素和环境相关因素的支持证据较少，因此对这两类因素的探索并不充分。Schaefer 等人(2014)基于人与自动化交互研究的元分析结果对三因素模型进行了修订，并对每个因素的内容都进行了更细致的划分。Lewis 等人(2022)则进一步发展了信任模型，其整合模型考虑了代理信任的影响，强调信任的动态调整过程，为不同类型的信任关系研究提供了一个通用的分析框架。可以看到过往信任模型的发展呈现出从静态到动态、从单纯的维度划分到更全面、细致的因素考量的趋势，但是仍存在忽视人与 AI 双向互信关系的局限。

3 人与 AI 动态互信的新模型

3.1 人与 AI 动态互信模型的提出

在通用人工智能时代背景下，人与 AI 的互动关系日趋复杂。过往人机信任模型，尽管在理论上有所贡献，但在解释人与 AI 之间动态且双向的信任关系方面存在局限，已不足以全面描述人与 AI 之间的信任交互过程。因此，本文拟提出一个新模型，旨在填补现有人与 AI 信任领域理论模型的空白。该模型充分参考已有信任模型的内容，尽量全面地把握影响信任过程的因素。在模型框架上，充分参考人际信任模型(Mcknight & Chervany, 1996)，包含委托人相关因素、受托人相关因素以及情境因素。每一类因素的具体内容参考通用性强的整合模型(Lewis & Marsh, 2022)，并进一步考虑人与 AI 互动的独特性。因此，新模型的特点体现在：模型强调信任不仅是人对 AI 的单向评估，而是一个涉及人和 AI 双方的互动过程，人和 AI 均会根据对方的行动和反馈，不断调整自身的信任水平和行为策略。综上，本文在已有的信任模型(包括人际信任模型、人机信任的四因素模型、人机信任的三因素模型、以及人对 AI 信任的整合模型)的基础上，针对通用人工智能时代人与 AI 双向互信的新型交互关系，提出了一个新的人机互信模型：人与 AI 动态互信模型，如图 1 所示。

该模型提出了人与 AI 信任中的两个重要特征：“互信”与“动态”。“互信”注重关系维度，而“动态”关注的是人与 AI 信任关系中的时程维度。

人与 AI 的互信关系是通用人工智能时代的新型人机关系，不同于以往研究中所关注的人对 AI 的单向信任，“互信”更加强调了 AI 在信任关系中与人相似的主体地位。随着通用人工智能技术的不断发展，智能机器将从一种支持人类操作

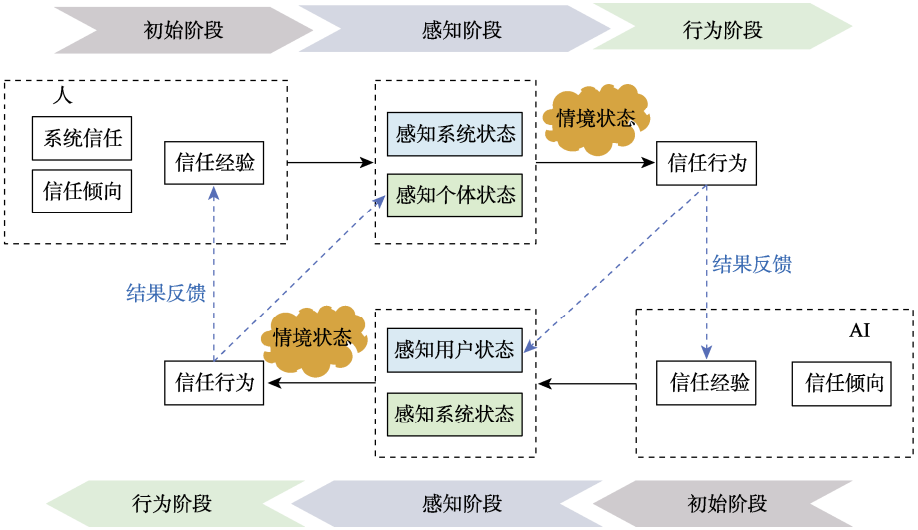

图 1 人与 AI 动态互信模型。信任方和受信方的角色是动态变化的，人与 AI 的信任受到感知对方状态(蓝色)、自身状态(绿色)和情境因素(黄色)的影响，并会根据结果反馈进行调整(蓝线)。彩图见电子版。

的辅助工具发展成为一个具有一定认知、独立执行、自适应等能力的自主化智能体(intelligent agent), 并且在一定程度上具备类似于人类的行为能力(Rahwan et al., 2019)。AI 将主动感知人类用户的状态, 和系统自身状态, 并由此评估对人类用户的信任水平, 决定控制权的归属。虽然在现实社会中尚未有相关实例, 但在科幻作品中已经描述了 AI 不相信用户从而拒绝用户使用工具的场景(Wikipedia contributors, 2024)。此时, 人机信任将不再是单向的人对机器系统的信任, 而是逐步转化为双向的, 即人机互信(许为, 葛列众, 2020; 许为 等, 2024)。人与 AI 互信实际上是基于人际信任的视角, 将人与 AI 视为对等的信任建立方。因此, 人与 AI 均可担任信任方(委托方)或被信方(受托方)的角色。

人与 AI 的互信关系也决定着其“动态”变化与以往单向信任在时程维度有所不同。在单向信任中, 按照信任发生、发展的时间顺序, 研究者将人与自动化系统的信任划分为倾向信任(Dispositional trust)、情境信任(Situational trust)和习得信任等几个阶段(Learned trust) (Hoff & Bashir, 2015; 高在峰 等, 2021; French et al., 2018; Merritt & Ilgen, 2008)。而在人际信任的建立过程中, 研究们提出了信任是一种反馈循环(feedback loop of trust), 信任方基于自身经验和倾向形成初始信任, 并基于对被信方的感知形成信任决策和行为, 之后根据反馈结果影响之后的信任(赵竞 等, 2013; Urban et al., 2009)。人与 AI 互信的交互过程是信任方和被信方根据信任过程中对方的状态和行为以及最终的信任结果持续调整自己的行为, 不断校准对被信方的信任水平的动态过程。综上, 本框架提出人与 AI 的动态互信可划分为三个阶段: 人与 AI 交互前的初始阶段、人与 AI 交互中的感知阶段以及行为阶段, 并且这三个阶段形成闭环。初始阶段是人与 AI 信任的最初阶段, 人与 AI 尚未接触, 依赖于自身固有的信任倾向、系统信任和以往交互中得到的相关信任经验等, 为之后的信任奠定基调。其中, 信任经验会在接收到本轮交互的结果反馈后得到矫正, 参与人机互信的动态过程; 而系统信任和信任倾向相对稳定, 不会参与后续的动态过程。在感知阶段, 人与 AI 的信任受到感知对方状态、感知自身状态和情境状态的影响, 形成信任决策。在行为阶段, 信任方完成信任行为, 并会根据行为结果反馈对初始阶段的信任经验进行校正, 从而产生新的信任经验, 同时对被信方的感知状态进行更新, 影响之后的信任行为。其中, 结果反馈包含了两层含义: 一方面是被信方的信任行为本身, 即被信方是否执行了信任方的决策; 另一方面是被信方执行或未执行信任方的决策之后所致系统运行的结果。人与 AI 互信通过以上过程不断校正, 实现信任的动态交互。比如当人作为被信方时, 如果 AI 传达出不信任的信号(比如疲劳状态警告), 人就会调整自身状态(相信 AI 的决策)以重获 AI 信任或者选择信任 AI 让其接管系统; 如果 AI 接收到人的不信任的指示, 也会通过系统自检(相信人类的决策)或者让设计者调试系统的方式调整自己的系统状态以争取获得人的信任, 从而达到系统正常运行的目的。人与 AI 互信的动态交互过程实际上反映的是信任校准过程(Lee & See, 2004)。虽然人与 AI 互信的理想状态是适当信任, 但实际上, 过度信任(高在峰 等, 2021; Robinette et al., 2016)与信任不足(Bigman & Gray, 2018; Longoni et al., 2019)在人机交互中十分常见。因此, 本文提出的模型认为, 人与 AI 的互信应该与人际信任相似, 存在信任更新(Kim et al., 2020; Mende-Siedlecki et al., 2013)。在人与 AI 的动态互信中, 信任更新取决于先前信任行为的结果。

综上, 人与 AI 的动态互信模型包含三个阶段(初始阶段、感知阶段和行为阶段)和两个主体(人与 AI)。在人与 AI 互信中的两个主体——人和 AI, 在前两个阶段分别存在相似和不同的信任影响因素, 下文将分别展开论述。

3.2　人对 AI 信任的影响因素

人对 AI 信任的影响因素主要是基于 Lewis 和 Marsh (2022)的整合模型框架结合过往文献提出。初始阶段, 人对 AI 的信任主要受到个体的信任倾向、过往的信任经验以及系统信任的影响。其中, 个体的信任倾向, 在其他研究中也被称为倾向性信任(Dispositional trust), 受到个体的固有特质的影响, 如年龄(Ma et al., 2020; Scopelliti et al., 2005)、人格(Rossi et al., 2018)、受教育程度(Liao & MacDonald, 2021)。信任经验, 是指通过使用人工智能相关的系统、产品而获得的先验经验或专业知识。这些经验有助于个人预测系统行为(Oleson et al., 2011), 从而改变人对 AI 的信任水平。比如,

Dikmen 和 Burns (2017)通过实验测试了用户对特斯拉汽车上自动驾驶系统的信任。研究结果表明, 那些经历过车辆意外事故的司机对自动驾驶系统的信任度较低; 相反, 对特斯拉自动驾驶系统有过了解的司机会对自动驾驶更加信任。人对 AI 的信任还受到系统信任(Institutional Trust)的影响。例如, 一些社会文化比其他社会文化更倾向于培养个体之间的普遍信任(Luhmann, 1990)。

在感知阶段, 人对 AI 的信任受到三方面因素的影响。一是感知个体状态, 即个体觉察自身是否能够胜任当前任务。二是感知系统状态, 包括感知可信和感知风险。其中感知可信包括对被信方的能力、可预测性、正直、仁慈以及代理信任(如品牌)等多维度的感知(Chen et al., 1995; Hoff & Bashir, 2015)。感知风险, 是指对被信方脆弱性和完成当前任务所伴随的风险水平的评估(Ajenaghughrure et al., 2020; Ma et al., 2020)。三是情境状态, 人需要对所处情境的性质和任务难度等进行评估。有研究表明, 当人机合作的任务工作量增加, 人对人工智能系统的信任度会降低, 人们会更倾向于独自完成任务(Oleson et al., 2011)。另一方面, Atoyan 等人(2006)通过实验证明, 当人机合作的任务过于复杂且繁多时, 人依靠自身能力无法完成任务, 可能会对合作的人工智能系统产生过度信任。

3.3 AI 对人信任的影响因素

在初始阶段, AI 对人信任的影响因素包括 AI 系统自身的信任倾向以及 AI 过往与用户交互过程中所形成的先验经验。其中, AI 的信任倾向目前主要是系统设计者对人类用户的信任倾向。考虑到目前仍然缺乏国家层面的人工智能相关法律制度(何积丰, 2019), 相关责任难以划分, 目前在高风险任务中 AI 往往倾向于信任人类用户(如自动驾驶)。在通用人工智能时代, AI 的信任倾向可能与人类相仿, 更多取决于 AI 的固有特质(如针对特定用户群体的个性化设计、主要任务、形态、安全保障等), 而非仅取决于初始设置。

在感知阶段, AI 对人的信任同样受三方面因素的影响。一是用户状态, AI 需要构建监测系统对使用者的状态(认知、生理、意图、情感、价值观、道德水平等)进行实时监测, 当使用者处于不可信任状态时(如疲劳、分心)AI 会主动接管以避免事故(许为 等, 2024)。二是系统状态, AI 需要对自身状态有一个主动监测和评估系统, 一方面是监测自身的性能和稳定性, 另一方面是评估当前状态是否能够完成任务。以自动驾驶为例, 自动驾驶汽车会配备大量的内部传感器, 以随时监测汽车内部状态数据, 并且研究者还在不断开发有效的自动故障诊断和健康监测算法(Biddle & Fallah, 2021)以评估系统状态。当系统监测到自身并不可靠时(如系统故障、任务超出系统能力), 就会做出信任人类的判断, 并提示人类用户接管控制权。三是情境状态, AI 需要对所处情境的风险程度、复杂程度进行评估, 比如环境状况、紧急情况的发生等, 以判断是否应该信任使用者。同样以自动驾驶为例, 汽车会使用摄像头、激光雷达、超声波传感器等传感器来感知交通路况、光照条件、障碍物情况等外部情境(Ignatious et al., 2022), 并根据感知到的情境采取相应的信任行为。当系统监测到高风险情境时(如汽车驾驶员即将发生追尾), AI 可能会更加谨慎, 减少对人类的信任, 采取刹车、紧急变道等紧急措施; 而在低风险情境下, AI 就会更信任人类, 给人类更多自主行为的权力。

4 人与 AI 信任的研究展望

目前, 通用人工智能时代即将拉开序幕, 然而, 有关人与 AI 信任的研究尚不充分, 未来研究可从以下三个方面展开:

4.1 AI 对人的信任

目前, AI 系统的设计者对于人类操作者的信任十分微妙。在一项针对优步司机的研究中, Möhlmann 和 Zalmanson (2017)指出, 持续的个人绩效评估和反馈(只有通过持续跟踪才能实现)违反了司机的自主意识, 降低了他们的信任。这种持续监控被视为一种微观管理方式, 说明部署人工智能的人(AI 系统的设计者)对于 AI 的操作者缺乏信任, 这反过来也导致司机对于自动驾驶 AI 信任度的降低。

智能时代, 机器逐渐具备类似人类的行为能力(Rahwan et al., 2019), 这时, 信任将不仅是单向的人对自动系统的信任, 而是双向的人机互信(许为, 葛列众, 2020)。在人机互信框架下, 以下两个方面的研究亟需开展。第一, 系统设计者对用户的信任。由于人类操作员的局限性(比如生理、心理), 在一些场景, 系统设计者往往更加信

任 AI 的判断, 而疏忽了对于用户状态的监测。比如, 研究者可基于驾驶员当前状态(疲劳、分心等)、系统状态(可靠性等)和情境状况(环境风险等)等数据进行建模, 构建系统对驾驶员的适当信任模型, 系统在驾驶员处于不可信状态时主动介入以避免事故。第二, 系统对用户的信任。如果说弱人工智能时代, AI 对人的信任可以等同为 AI 设计者对于用户的信任, 那么在强人工智能时代, AI 将具备自我意识与自主判断, AI 与人的合作关系将取决于双方互信。一个崭新的问题则是人类如何赢得 AI 的信任, 是否如同人−人信任一样, 其特殊性又将体现在哪些方面? 随着 AI 智能水平的提高, 这一问题将逐步影响人与 AI 的交互。

4.2　人与 AI 互信的量化模型

目前, 大多数研究从理论上提出了一些研究框架和定性的模型, 尚缺乏人与 AI 互信的量化模型指导系统设计。量化模型的建立, 取决于以下两个先决要素。

(1)信任的测量。目前最常用的信任测量方法是自我报告法(Fogg & Tseng, 1999; Jian et al., 2000; Madsen & Gregor, 2000)。自我报告测量方法易于使用, 如果研究者正确构建了问卷或量表, 那么该方法可以有效地反映操作者的人机信任水平。然而, 自我报告的测量方法对交互任务具有干扰性并且难以实时捕获人与 AI 信任的动态变化, 它在实际环境中的应用受到很大限制。此外, 该方法具有不可避免的缺陷, 即被试可能不能或不愿意准确报告他们的真实态度, 并且他们无法描述隐性态度对其信任水平的影响(Stokes et al., 2010)。为了弥补自我报告测量的缺陷, 一些研究者开始从可见的行为中来推断人机信任水平。使用行为指标度量人机信任主要是依据遵从和依赖的概念, 即当操作员更倾向于遵从或依赖系统时, 其人机信任水平较高, 反之则较低。遵从是指当机器系统发出信号时, 操作者做出响应, 可以利用操作者对系统所提供建议或动作的接受程度进行测量(Bindewald et al., 2018); 依赖则是指当机器系统处于沉默状态或正常运行状态时, 操作者不响应, 可以用操作者使用自动化系统的时间(次数)占总时间(总任务次数)的比例进行测量(de Vries et al., 2003; Gremillion et al., 2016)。此外还有使用反应时进行行为测量的方法, 可以通过操作者察觉到系统风险后, 接管自动化系统控制权的速度进行测量(Molnar et al., 2018; Payre et al., 2016), 操作者的反应越快, 则表示越不信任自动化系统。生理及神经测量旨在通过测量与人机信任相关的生理及神经指标来对人机信任进行实时测量, 虽然该方法尚处于起步阶段, 但已有文献表明它在获取人机信任的实时动态变化方面非常有效(Akash et al., 2018)。

在现有测量方法的基础上, 仍需进一步识别多种测量方法结果之间的不一致, 寻找行为指标和生理及神经指标与主观信任水平的对应关系, 确定更加准确的实时测量指标, 以动态识别人与 AI 信任的基本状态(适当的信任、信任不足和过度信任)。如何准确表征 AI 对人的信任, 也将成为人与 AI 信任研究中的重要内容。

(2)信任影响因素的权重确定。不同的研究针对人与 AI 互信中的各个影响因素展开了相关研究, 然而, 目前尚缺乏整合模型的相关研究。如何准确测量各种信任相关的影响因素, 并在实际的人与 AI 互动的过程中, 分析和量化各因素的权重、作用条件, 涵盖产生重要影响的环境及个体因素, 并且考虑未建模因素对模型性能的不良影响, 构建满足不同设计阶段需求的信任计算模型将成为未来改进人与 AI 互信, 提高模型应用价值的关键。

4.3　多智能体互动中的人与 AI 互信

本文提出的模型适用于 AI 作为人类助手或者协作伙伴、人机协作完成任务等常见情境(Mohanty & Vyas, 2018), 但是模型仅关注了单一人类与单个 AI 互动时的互信过程, 随着 AI 使用场景的复杂化, 将会涉及到多个人类与多个 AI 之间的互动。以往研究者认为, 在多智能体互动中, 每个成员所担任的角色以及成员之间互动的方式都是影响信任的关键因素(Yagoda & Gillan, 2012)。在这样的环境中, 信任的动态构建过程变得更加复杂。可以在本模型的基础上进一步纳入各智能体的身份角色, 考虑其在动态互信过程中的权重。举例而言, 图 2 在分布式认知(Perry, 2003)的基础上, 纳入了人与 AI 动态互信过程中角色的分配。当多智能体互动中出现“意见领袖”时(图中蓝色智能体), 意见领袖(可能是人类或 AI)的信任经验将通过交流, 进而影响到其他智能体(图中灰色智能体)。只有将人工智能放到复杂群体(如团队或网络)中进行研究, 研究人员才能真正理解人们

与人工智能建立“合作伙伴”关系的方式，以及人工智能如何改变人与人之间以及人与其他机器之间关系的方式。此外，由于人工智能的行为不是稳定不变的，学者们需要研究它基于人类与人工智能交互的变化方式(Rahwan et al., 2019)，以促进对关系变化的了解。未来的研究应该考虑多人与多个 AI 的交互，这将为建立人与 AI 的伙伴关系，形成人在回路(human-in-the-loop)框架下的人与 AI 互信提供更好的支持。

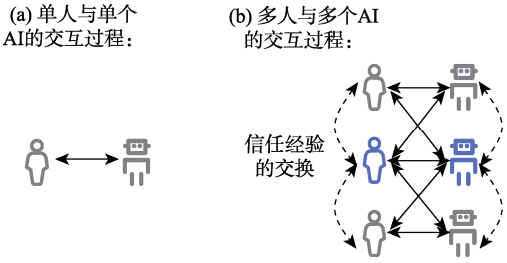

图 2 人与 AI 的信任交互过程。a) 单人与单个 AI 交互(实线), b)多人与多个 AI 的交互过程。其中，人/AI 交互(实线)集中的节点即意见领袖(蓝色)，意见领袖可能在不同的主体(灰色)之间形成信任经验的交换(虚线)，进而影响到其他主体之间的信任。彩图见电子版。

参考文献

高在峰, 李文敏, 梁佳文, 潘晗希, 许为, 沈模卫. (2021). 自动驾驶车中的人机信任. *心理科学进展*, *29*(12), 2172−2183.

何积丰. (2019). 安全可信人工智能. *信息安全与通信保密*, *10*, 5−8.

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.

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.

闫宏秀. (2019). 用信任解码人工智能伦理. *人工智能*, (4), 7.

赵竞, 孙晓军, 周宗奎, 魏华, 牛更枫. (2013). 网络交往中的人际信任. *心理科学进展*, 21(8), 1493−1501.

Ajenaghughrure, I. B., da Costa Sousa, S. C., & Lamas, D. (2020, June). Risk and trust in artificial intelligence technologies: A case study of autonomous vehicles. In *2020 13th International Conference on Human System Interaction* (pp. 118−123), Tokyo, Japan. doi: 10.1109/HSI49210.2020.9142686

Akash, K., Hu, W.-L., Jain, N., & Reid, T. (2018). A classification model for sensing human trust in machines using EEG and GSR. *ACM Transactions on Interactive Intelligent Systems*, *8*(4), 1−20. doi: 10.1145/3132743

Aly, A., & Tapus, A. (2016). Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human−robot interaction. *Autonomous Robots*, *40*(2), 193−209. doi: 10.1007/s10514-015-9444-1

Atoyan, H., Duquet, J.-R., & Robert, J.-M. (2006, April). Trust in new decision aid systems. In *Proceedings of the 18th Conference on I'Interaction Homme-Machine* (pp. 115−122), Montreal, Canada. doi: 10.1145/1132736.1132751

Bartneck, C., & Forlizzi, J. (2004, September). A design-centered framework for social human-robot interaction. In *IEEE International Workshop on Robot & Human Interactive Communication* (pp. 591−594), Kurashiki, Japan. doi: 10.1109/ROMAN.2004.1374827

Biddle, L., & Fallah, S. (2021). A novel fault detection, identification and prediction approach for autonomous vehicle controllers using SVM. *Automotive Innovation*, *4*(3), 301−314. doi: 10.1007/s42154-021-00138-0

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21−34. doi: 10.1016/j.cognition.2018.08.003

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

Bindewald, J. M., Rusnock, C. F., & Miller, M. E. (2018). Measuring human trust behavior in human-machine teams. In *Advances in Human Factors in Simulation and Modeling* (Vol. 591, pp. 47−58), Los Angeles, USA. Springer International Publishing. doi: 10.1007/978-3-319-60591-3_5

Binz, M. & Eric Schulz. (2023). Using cognitive psychology to understand GPT-3. Proceedings of the National Academy of Sciences, 120(6), e2218523120. doi: 10.1073/pnas.2218523120

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* preprint arxiv: 2303. 12712.

Chen, I.-R., Bastani, F. B., & Tsao, T.-W. (1995). On the reliability of AI planning software in real-time applications. *IEEE Transactions on Knowledge and Data Engineering*, *7*(1), 4−13. doi: 10.1109/69.368522

Chen, J. Y. C., Barnes, M. J., & Harper-Sciarini, M. (2011). Supervisory control of multiple robots: Human-performance issues and user-interface design. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *41*(4), 435−454. doi: 10.1109/TSMCC.2010.2056682

Christoforakos, L., Gallucci, A., Surmava-Große, T., Ullrich, D., & Diefenbach, S. (2021). Can robots earn our trust the same way humans do? A systematic exploration of competence, warmth, and anthropomorphism as determinants of trust development in HRI. *Frontiers in Robotics and AI*, *8*, 640444. doi: 10.3389/frobt.2021.

640444

Cofta, P. (2007). *Trust, complexity and control: Confidence in a convergent world*. John Wiley & Sons, Ltd. doi: 10.1002/9780470517857

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331−349. doi: 10.1037/xap0000092

de Vries, P., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies*, *58*(6), 719−735. doi: 10.1016/S1071-5819(03)00039-9

Deutsch, M. (1962). Cooperation and trust: Some theoretical notes. In Jones, M.R., (Ed.), *Nebraska symposium on motivation* (pp. 275−320). University of Nebraska Press.

Dikmen, M., & Burns, C. (2017, October). Trust in autonomous vehicles: The case of Tesla Autopilot and Summon. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 1093−1098), Banff, Canada. doi: 10.1109/SMC.2017.8122757

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., … Sandvig, C. (2015, April). "I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 153−162), Seoul, Republic of Korea. doi: 10.1145/2702123.2702556

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. doi: 10.1037/0022-3514.82.6.878

Fiske, S. T., Xu, J., Cuddy, A. C., & Glick, P. (1999). (Dis)respecting versus (Dis)liking: Status and interdependence predict ambivalent stereotypes of competence and warmth. *Journal of Social Issues*, *55*(3), 473−489. doi: 10.1111/0022-4537.00128

Fogg, B. J., & Tseng, H. (1999, May). The elements of computer credibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 80−87), Pittsburgh, USA. doi: 10.1145/302979.303001

Forcier, M. B., Khoury, L., & N Vézina. (2020). Liability issues for the use of artificial intelligence in health care in canada: AI and medical decision-making. *Dalhousie Medical Journal, 46*(2), 7−11. doi: 10.15273/dmj.Vol46No2.10140

French, B., Duenser, A., & Heathcote, A. (2018). Trust in automation − A literature review. *Commonwealth Scientific and Industrial Research Organisation Report,* EP184082.

Frison, A.-K., Wintersberger, P., Riener, A., Schartmüller, C., Boyle, L. N., Miller, E., & Weigl, K. (2019, May). In UX we trust: Investigation of aesthetics and usability of driver-vehicle interfaces and their impact on the perception of automated driving. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1−13), Glasgow, UK. doi: 10.1145/3290605.3300374

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627−660. doi: 10.5465/annals.2018.0057

Gockley, R., Simmons, R., & Forlizzi, J. (2006, September). Modeling affect in socially interactive robots. In *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 558−563), Hatfield, UK. doi: 10.1109/ROMAN.2006.314448

Gremillion, G. M., Metcalfe, J. S., Marathe, A. R., Paul, V. J., Christensen, J., Drnec, K., … Atwater, C. (2016). Analysis of trust in autonomy for convoy operations. In *Micro and nanotechnology sensors, systems, and applications, 9836*, 356−365. doi: 10.1117/12.2224009

Groom, V., & Nass, C. (2007). Can robots be teammates?: Benchmarks in human−robot teams. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, *8*(3), 483−500. doi: 10.1075/is.8.3.10gro

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(5), 517−527. doi: 10.1177/0018720811417254

Hancock, P. A., Nourbakhsh, I., & Stewart, J. (2019). On the future of transportation in an era of automated and autonomous vehicles. *Proceedings of the National Academy of Sciences, 116*(16), 7684−7691. doi: 10.1073/pnas.1805770115

Hardin, R. (2002). *Trust and trustworthiness*. Russell Sage Foundation.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407−434. doi: 10.1177/0018720814547570

Ignatious, H. A., & Khan, M. (2022). An overview of sensors in autonomous vehicles. *Procedia Computer Science*, *198*, 736−741. doi: 10.1016/j.procs.2021.12.315

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. doi: 10.1207/S15327566IJCE0401_04

Khastgir, S., Birrell, S., Dhadyalla, G., & Jennings, P. (2017). Calibrating trust to increase the use of automated systems in a vehicle. In *Advances in Human Aspects of*

Transportation: Proceedings of the AHFE 2016 International Conference on Human Factors in Transportation, *484*, 535−546. Springer International Publishing. doi: 10.1007/978-3-319-41682-3_45

Kim, M., Park, B. K., & Young, L. (2020). The psychology of motivated versus rational impression updating. *Trends in Cognitive Sciences*, *24*(2), 101−111. doi: 10.1016/j.tics.2019.12.001

Kulms, P., & Kopp, S. (2018). A social cognition perspective on human−computer trust: The effect of perceived warmth and competence on trust in decision-making with computers. *Frontiers in Digital Humanities*, *5*, 14. doi: 10.3389/fdigh.2018.00014

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50−80. doi: 10.1518/hfes.46.1.50_30392

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. doi: 10.1016/j.cogsys.2021.11.001

Liao, T., & MacDonald, E. F. (2021). Manipulating users' trust of autonomous products with affective priming. *Journal of Mechanical Design*, *143*(5), 051402. doi: 10.1115/1.4048640

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629−650. doi: 10.1093/jcr/ucz013

Luhmann, N. (1990). Technology, environment and social risk: A systems perspective. *Organization & Environment, 4*(3), 223−231. doi: 10.1177/108602669000400305

Ma, Y., Li, S., Qin, S., & Qi, Y. (2020, December). Factors affecting trust in the autonomous vehicle: A survey of primary school students and parent perceptions. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications* (pp. 2020−2027), Guangzhou, China. doi: 10.1109/TrustCom50675.2020.00277

Madsen, M., & Gregor, S. (2000, December). Measuring human-computer trust. In *11th Australasian Conference on Information Systems* (Vol. 53, pp. 6−8), Brisbane, Australia.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709−734. doi: 10.5465/amr.1995.9508080335

Mcknight, D. H., & Chervany, N. L. (1996). *The meaning of trust* [Technical Report]. Management Information Systems Research Center, University of Minnesota.

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience*, *8*(6), 623−631. doi: 10.1093/scan/nss040

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in humanautomation interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(2), 194−210. doi: 10.1518/001872008X288574

Mohanty, S., & Vyas, S. (2018). Putting it all together: Toward a human-machine collaborative ecosystem. In S. Mohanty & S. Vyas (Eds.), *How to Compete in the Age of Artificial Intelligence: Implementing a collaborative human-machine strategy for your business* (pp. 215−229), Apress, Berkeley, CA, USA. doi: 10.1007/978-1-4842-3808-0_11

Möhlmann, M., & Zalmanson, L. (2017, December). Hands on the wheel: Navigating algorithmic management and Uber drivers'. In *Autonomy', in proceedings of the international conference on information systems* (pp. 10−13), Seoul, Republic of Korea.

Molnar, L. J., Ryan, L. H., Pradhan, A. K., Eby, D. W., St. Louis, R. M., & Zakrajsek, J. S. (2018). Understanding trust and acceptance of automated vehicles: An exploratory simulator study of transfer of control between automated and manual driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *58*, 319−328. doi: 10.1016/j.trf.2018.06.004

Noah, B. E., Gable, T. M., Schuett, J. H., & Walker, B. N. (2016, October). Forecasted affect towards automated and warning safety features. In *Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 123−128), Ann Arbor, USA. doi: 10.1145/3004323.3004337

Noah, B. E., Wintersberger, P., Mirnig, A. G., Thakkar, S., Yan, F., Gable, T. M., Kraus, J., & McCall, R. (2017, September). First workshop on trust in the age of automated driving. *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications Adjunct* (pp. 15−21), Oldenburg, Germany. doi: 10.1145/3131726.3131733

Oleson, K. E., Billings, D. R., Kocsis, V., Chen, J. Y. C., & Hancock, P. A. (2011, February). Antecedents of trust in human-robot collaborations. In *2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)* (pp. 175−178), Miami Beach, USA. doi: 10.1109/COGSIMA.2011.5753439

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230−253. doi: 10.1518/001872097778543886

Payre, W., Cestac, J., & Delhomme, P. (2016). Fully automated driving: Impact of trust and practice on manual control recovery. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(2), 229−241. doi: 10.1177/0018720815612319

Perry, M. (2003). Distributed cognition. In J.M. Carroll (Ed.), *HCI models, theories, and frameworks: Toward a*

multidisciplinary science (pp. 193−223), Morgan Kaufmann.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477−486. doi: 10.1038/s41586-019-1138-y

Raj, M., & Seamans, R. (2019). Primer on artificial intelligence and robotics. *Journal of Organization Design*, *8*(1), 11. doi: 10.1186/s41469-019-0050-0

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016, March). Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 101−108), Christchurch, New Zealand. doi: 10.1109/HRI.2016.7451740

Rödel, C., Stadler, S., Meschtscherjakov, A., & Tscheligi, M. (2014, September). Towards autonomous cars: The effect of autonomy levels on acceptance and user experience. *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 1−8), Seattle, USA. doi: 10.1145/2667317.2667330

Rossi, A., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2018). The impact of peoples' personal dispositions and personalities on their trust of robots in an emergency scenario. *Paladyn, Journal of Behavioral Robotics*, *9*(1), 137−154. doi: 10.1515/pjbr-2018-0010

Sanders, T., Oleson, K. E., Billings, D. R., Chen, J. Y. C., & Hancock, P. A. (2011). A model of human-robot trust: Theoretical model development. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*(1), 1432−1436. doi: 10.1177/1071181311551298

Schaefer, K. E., Billings, D. R., Szalma, J. L., Adams, J. K., Sanders, T. L., Chen, J. Y., & Hancock, P. A. (2014). *A meta-analysis of factors influencing the development of trust in automation: Implications for human-robot interaction* [Technical Report]. Army Research Lab, Aberdeen Proving Ground, Maryland, Human Research Engineering Directorate. doi: 10.21236/ADA607926

Scopelliti, M., Giuliani, M. V., & Fornara, F. (2005). Robots in a domestic setting: A psychological approach. *Universal Access in the Information Society*, *4*(2), 146−155. doi: 10.1007/s10209-005-0118-1

Shiffrin, R., & Mitchell, M. (2023). Probing the psychology of AI models. *Proceedings of the National Academy of Sciences*, *120*(10), e2300963120.

Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) ethics: Ethics of AI and ethical AI. *Journal of Database Management*, *31*(2), 74−87. doi: 10.4018/JDM.2020040105

Stephanidis, C., Salvendy, G., Antona, M., Chen, J. Y. C., Dong, J., Duffy, V. G., … Zhou, J. (2019). Seven HCI grand challenges. *International Journal of Human−Computer Interaction*, *35*(14), 1229−1269. doi: 10.1080/10447318.2019.1619259

Stokes, C. K., Lyons, J. B., Littlejohn, K., Natarian, J., Case, E., & Speranza, N. (2010, May). Accounting for the human in cyberspace: Effects of mood on trust in automation. In *2010 International Symposium on Collaborative Technologies and Systems* (pp. 180−187), Chicago, USA. doi: 10.1109/CTS.2010.5478512

Sullins, J. P. (2010). Love and sex with robots: The evolution of human-robot relationships [Book review]. *Industrial Robot*, *37*(4), 401−402. doi: 10.1108/ir.2010.04937dae.001

Urban, G. L., Amyx, C., & Lorenzon, A. (2009). Online trust: State of the art, new frontiers, and research potential. *Journal of Interactive Marketing*, *23*(2), 179−190. doi: 10.1016/j.intmar.2009.03.001

van Pinxteren, M. M. E., Wetzels, R. W. H., Rüger, J., Pluymaekers, M., & Wetzels, M. (2019). Trust in humanoid robots: Implications for services marketing. *Journal of Services Marketing*, *33*(4), 507−518. doi: 10.1108/JSM-01-2018-0045

Walter, S., Wendt, C., Böhnke, J., Crawcour, S., Tan, J.-W., Chan, A., … Traue, H. C. (2014). Similarities and differences of emotions in human−machine and human−human interactions: What kind of emotions are relevant for future companion systems? *Ergonomics*, *57*(3), 374−386. doi: 10.1080/00140139.2013.822566

Wang, W., & Siau, K. (2019). Artificial Intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda. *Journal of Database Management*, *30*(1), 61−79. doi: 10.4018/JDM.2019010104

Wikipedia contributors. (2024, January 13). Psycho-Pass. In *Wikipedia, The Free Encyclopedia*. from https://en.wikipedia.org/w/index.php?title=Psycho-Pass&oldid=1195338833

Wintersberger, P., Noah, B. E., Kraus, J., McCall, R., Mirnig, A. G., Kunze, A., … Walker, B. N. (2018, September). Second workshop on trust in the age of automated driving. *Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 56−64), Toronto, Canada. doi: 10.1145/3239092.3239099

Wright, P., McCarthy, J., & Meekison, L. (2003). Making sense of experience. In Blythe, M.A., Overbeeke, K., Monk, A. F., Wright, P. C. (Eds.), *Funology: From usability to enjoyment* (Vol. 3, pp. 43−53), Springer Netherlands. doi: 10.1007/1-4020-2967-5_5

Yagoda, R. E., & Gillan, D. J. (2012). You want me to trust a robot? The development of a human−robot interaction trust scale. *International Journal of Social Robotics*, *4*(3), 235−248. doi: 10.1007/s12369-012-0144-0

Human-AI mutual trust in the era of artificial general intelligence

QI Yue[1,2], CHEN Junting[1,2], QIN Shaotian[1,2], DU Feng[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] *CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing* 100101, *China*)

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

Abstract: With the development of technology, artificial general intelligence has begun to take shape, ushering in a new era for human-machine interaction and relationships. The trust between humans and artificial intelligence (AI) are on the brink of a transformative shift from unidirectional trust, where people trust AI, to a state of mutual trust between humans and AI. This study, based on a review of the interpersonal trust model in social psychology and the human-machine trust model in engineering psychology, proposes a dynamic mutual trust model for human-AI relationships from the perspective of interpersonal trust. The model regards humans and AI as equal contributors to trust-building, highlighting the "mutual trust" in the relational dimension and the "dynamics" in the temporal dimension of trust between humans and AI. It constructs a fundamental framework for dynamic mutual trust between humans and AI, incorporating influencing factors, result feedback, and behavior adjustment as essential components. This model pioneers the inclusion of AI's trust towards humans and the dynamic interactive process of mutual trust, offering a new theoretical perspective for the study of trust between humans and AI. Future research should focus on understanding the establishment and maintenance of trust from AI towards humans, developing quantitative models for human-AI trust, and exploring mutual trust dynamics within multi-agent interactions.

Keywords: trust, human-machine mutual trust, trust calibration, human-machine relationship, human-AI