人机团队信任发展的基于智能体的仿真：经实证验证的框架

Ravi Kalluri职业研究学院 东北大学 –
圣何塞，加利福尼亚州，美国
r.kalluri@northeastern.edu

摘要

人机团队日益广泛地部署于复杂的作业环境中，但能够预测不同条件下团队绩效的已验证仿真工具仍然稀缺。本文提出一种以实证为基础的基于智能体的模型，用于刻画人机团队中的信任动态、工作负荷分配和协作绩效。该模型基于 NetLogo 6.4.0 实现，可模拟由 2–10 个智能体组成的团队执行不同复杂程度的任务。我们以 Hancock 等人（2021年）对人机交互研究的荟萃分析为基准验证这一方法，在 8 个信任前因类别中实现了 4 个类别的区间效度，并展现出较强的序数效度（斯皮尔曼 $\rho = 0.833$），表明模型能够正确再现相对效应大小。采用单因素逐次分析（OFAT）和全因子设计开展敏感性分析（每种条件下重复实验 = 50 次）发现，机器人可靠性对信任的影响最大（$\eta 2 = 0.35$），并且对任务成功（$\eta 2 = 0.93$）和生产力（$\eta 2 = 0.89$）的影响占主导地位，这与荟萃分析结果一致。与预期相反，各情景中的信任不对称比率介于 0.07 至 0.55 之间，低于 1.50 的荟萃分析基准，揭示了一项边界条件：当信任修复机制持续发挥作用时，单次事件的不对称并不必然带来累积不对称。情景分析发现了信任与绩效脱钩现象：信任恢复情景的生产力最高（4.29），尽管其信任程度最低（38.2）；而不可靠机器人情景的信任程度最高（73.2），尽管其任务成功率最低（33.4%）。这些发现表明，校准误差——主观信任与客观能力之间的差异——是一项有别于信任程度的关键诊断指标。析因方差分析证实，可靠性、透明度、沟通和协作均存在显著主效应（$p < .001$），合计解释了 45.4% 的信任方差，而交互效应极小（$\Sigma \eta 2 = 0.02$）。这项开源实现为从业者提供了一种以证据为基础的工具，用于探索人机团队配置，尤其有助于在部署前识别过度信任和信任不足的情形。

关键词：人机交互；团队动态；基于智能体的建模；信任校准；团队绩效；社会仿真

1. 引言

将机器人整合到人类工作团队中，代表着组织开展协作工作方式的根本转变。与人类从远处监督机器的传统自动化不同，现代人机协作需要持续交互、相互适应，并建立适当的信任关系（Hancock 等人，2021）。这一演变在制造业、医疗保健和搜救行动中尤为明显，并且越来越多地出现在协作机器人与人类操作员并肩工作的知识工作环境中。尽管人机团队的部署日益增多，从业者仍缺乏已验证的工具，无法在代价高昂的实现之前预测团队绩效。当前方法主要依赖试错，这可能导致团队失败、操作员信任受损以及潜在的安全事故。这一挑战因人机交互所特有的信任、性能、压力与沟通之间复杂的非线性关系而进一步加剧。

三项关键缺口限制了我们当前对人机团队的理解以及设计有效人机团队的能力。首先，尽管大量实证研究已考察影响对机器人信任的各个因素，但我们仍缺乏能够刻画这些因素如何随时间动态相互作用的整合模型。Hancock 等人（2021）的全面荟萃分析将机器人性能、机器人属性、人类特征和环境因素确定为信任前因，但尚无仿真框架系统地对照这些既定效应量进行验证。其次，信任发展的时间动态仍未得到充分理解。实验室研究通常在离散时间点考察信任，而现实世界中的团队所经历的信任演变是连续的，并受到累积互动、失败与恢复的塑造。第三，从业者需要关于参数阈值的可操作指导——即能够预测成功协作的可靠性、透明度和团队构成的具体数值。

本文通过一个基于智能体的模型来填补这些缺口，该模型整合了完整的 Hancock 等人（2021）三因素信任框架，并对照荟萃分析效应量验证了预测结果。我们的具体贡献包括：（1）一个以实证为基础的仿真框架，能够在其置信区间内复现已发表的信任前因相关系数；（2）展示了涌现的信任不对称性，这与“信任下降快于信任建立”这一既定发现相一致；（3）识别了对系统设计具有实际意义的关键参数阈值；（4）提供开源实现，以同时支持研究复现和从业者应用。

本文其余部分安排如下。第2节回顾理论基础和已有的建模方法。第3节介绍模型设计与框架。

第4节介绍实现细节。第5节报告与实证基准的验证结果。第6节介绍仿真场景及其研究发现。第7节讨论理论和实践意义，第8节总结局限性并展望未来方向。

2. 理论背景与相关工作

本节综合梳理人机交互中的信任研究，阐明指导模型开发的理论框架，并回顾先前的仿真方法。我们不逐一罗列研究发现，而是着重确立用于验证本模型的实证基准。

2.1 信任：人机协作的基础

信任从根本上决定了人类是否会在重要任务中依赖机器人队友。Lee和See（2004）将对自动化的信任定义为“在具有不确定性和脆弱性的情境中，相信某个智能体会帮助实现个人目标的态度”。该定义强调了两个关键方面：人类与机器人之间的目标一致性，以及人类在将任务委托给自动化系统时对脆弱性的接纳。

适当的信任校准——使信任水平与系统的实际能力相匹配——已成为人机交互研究的核心议题（Wischnewski等人，2023年）。信任不足会使操作员无法充分利用性能良好的系统，而过度信任则会导致不当依赖可能发生故障的系统（Parasuraman和Riley，1997；Robinette等人，2017）。这两种校准失调都会带来实际后果：信任不足会降低效率，过度信任则会损害安全。系统设计人员面临的挑战是创造有利于适当信任校准的条件，而这需要理解影响信任发展的因素。

2.1.1 人机交互中的信任

信任是影响人机团队效能的关键中介。Lee和See（2004）将信任定义为“在具有不确定性和脆弱性的情境中，智能体将帮助实现个体目标的态度”，并指出信任有三个依据：性能（系统做什么）、过程（系统如何运行）和目的（系统为何如此设计）。Schaefer et al.（2024）进一步拓展了这一框架，将信任描述为一种非对称关系：人类信任者基于感知可信度主动产生信任；这强调了可信度是机器人的一种属性，而信任是人类的一种态度（Hancock 等人，2023）。Mayer等人（1995）将能力、善意和正直确定为信任前因。

在应用于人机交互情境时，能力对应机器人的可靠性和专业知识，善意对应温暖感和沟通行为，而正直对应透明度和可预测性。Malle和Ullman（2021）进一步纳入道德维度，指出对机器人的信任既涉及能力评估，也涉及性格归因。Hoff和Bashir（2015）区分了倾向性信任（稳定的个体差异）、情境性信任（取决于情境的因素）和习得信任（基于经验的更新）。Loizaga等人（2024）为这一分层架构提供了实证支持，表明初始交互对信任建立至关重要，且倾向性信任和习得信任通过不同机制发挥作用。

2.1.2 信任校准

信任校准——即主观信任与实际系统能力之间的匹配——已成为区别于信任程度的一种独立构念（de Visser et al., 2020; Chiou & Lee, 2023; Lucas et al., 2024）。当信任超过系统能力所能支持的合理水平时，就会出现过度信任，可能导致误用和安全违规。当信任低于合理水平时，就会出现信任不足，导致弃用和低效率（Lee & See, 2004; Parasuraman & Riley, 1997）。De Visser et al. (2020) 提出了关系公平这一概念，即代表参与者之间累积善意的情感资源，并将其视为信任校准随时间变化的预测因素。他们的框架描述当前的信任动态，预测互动的影响，并提出促进校准的干预措施。Kox et al. (2024) 表明，透明度机制可以防止过度信任（通过设定现实预期）和信任不足（通过澄清意外行为）。

2.1.3 信任不对称性

心理学文献一再表明，与同等程度的正面事件相比，负面事件对态度的影响不成比例地更大（Baumeister 等人，2001；Rozin & Royzman，2001）。Slovic（1993）的不对称原则——信任十分脆弱，逐步建立，却会迅速瓦解——已在人机交互情境中得到验证。Hancock 等人（2021）的荟萃分析证实，信任不对称比约为 1.50，这表明失败导致的信任衰减超过成功带来的信任形成。然而，不对称效应可能取决于情境因素。Hopko 等人（2022）强调，信任动态会随着时间推移受到互动、背景和环境的影响。单次事件不对称性与累积信任结果之间的关系，可能受到事件频率和信任修复机制可用性的调节。

2.1.4 信任违背与修复

当机器人行为未能满足人类期望时，就会发生信任违背，并降低感知可信度（Esterwood & Robert，2022）。人际关系文献中提出的修复策略——道歉、否认、解释和承诺——已在人机交互情境中得到研究，但结果不一。Esterwood and Robert（2023a）表明，发生三次信任违背后，这些策略都无法完全恢复可信度，这支持了修复存在“三振出局”上限的观点。后续融合“心智理论”的研究（Esterwood & Robert, 2023b）显示，心智感知会调节修复效果：当人类认为机器人具备情绪能力时，道歉更有效；当人类认为机器人具有意向性时，解释更有效。Tsumura and Yamada（2024）进一步表明，在一系列任务中，共情型机器人行为有助于修复信任。

2.1.5 元分析框架

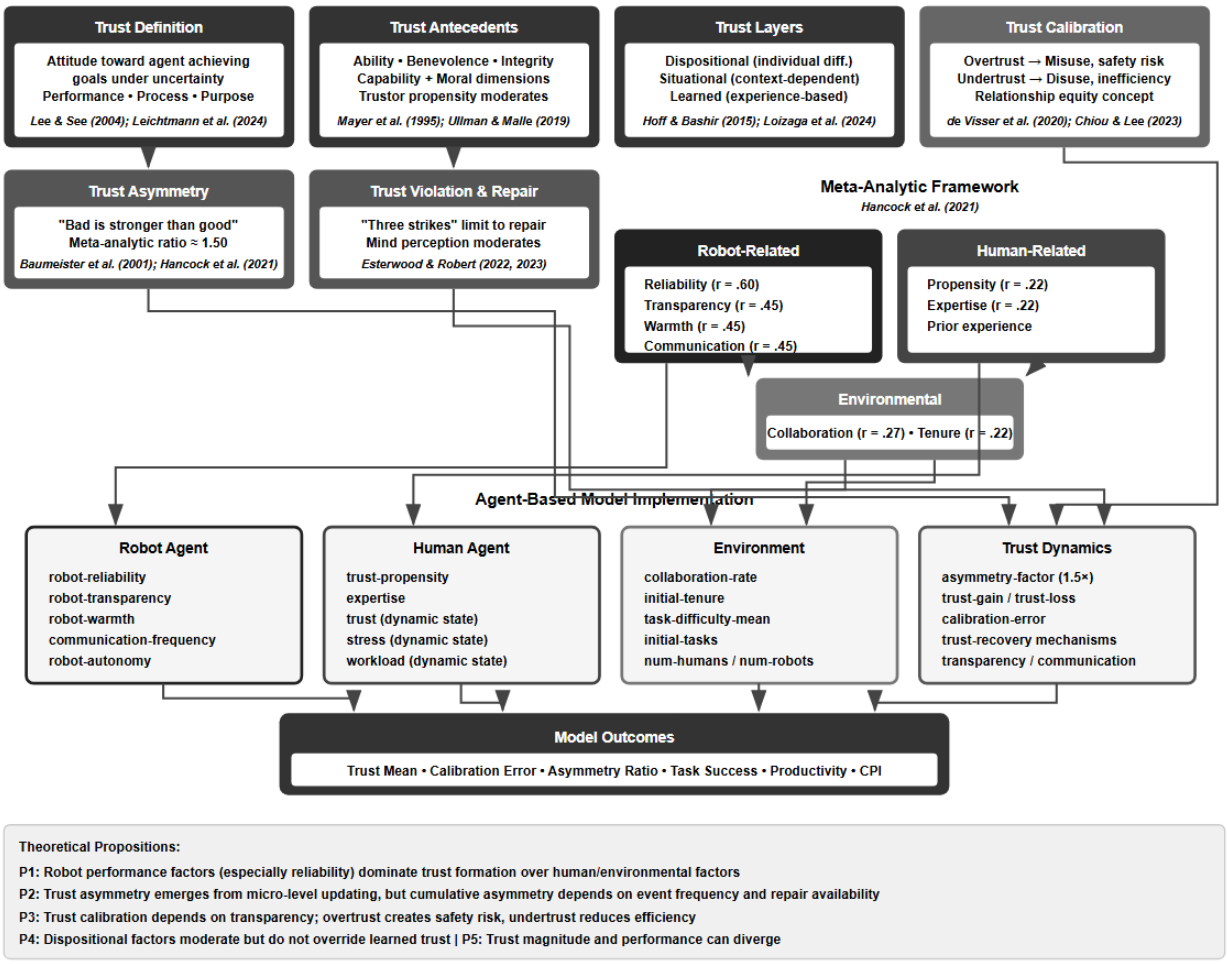
Hancock 等人（2021年）综合分析了 69 项研究（N = 7,769），其全面的荟萃分析为模型验证提供了实证基础。他们将信任前因归为三个因素：

机器人相关因素包括可靠性（r = .60，95%置信区间 [.54，.65]）、透明度（r = .45 [.37，.52]）、温暖感（r = .45 [.37，.52]）和沟通（r = .45 [.37，.52]）。机器人性能是信任最强的预测因素。

人为因素包括信任倾向（r = .22 [.14，.30]）和专业知识（r = .22 [.14，.30]），这表明信任者特征对信任的影响独立于机器人行为。

环境因素包括协作率（r = .27 [.19，.35]）和任职年限（r = .22 [.14，.30]）。Schaefer et al.（2024年）证实，这些因素类别的相对重要性因情境而异：机器人相关因素在工业场景中占主导地位，而人类因素和环境因素在社会照护应用中变得更加重要。

2.1.6 理论整合



Fig图 1. 人机团队信任动态的理论框架。

该仿真通过统一架构整合了这些理论视角（图1）。人类智能体具有倾向性属性（信任倾向、专业知识）和动态状态（当前信任、压力）。机器人智能体具有固定属性（可靠性、透明度、温暖感），这些属性会影响交互结果。任务分配会产生成功或失败的结果，并触发信任更新；不对称因子则调节积极体验和消极体验的相对影响。该模型通过追踪主观信任与客观能力之间的差异来实现校准，从而识别过度信任和信任不足的情况。信任修复动态将透明度和沟通纳入其中，作为违规发生后可在一定限度内恢复信任的机制。

2.1.7 理论命题

整合框架提出了四个可检验命题：

P1: 与人类相关因素或环境因素相比，机器人性能因素对信任的影响更大（Hancock 等人，2021年）。

P2：信任不对称性源于微观层面的更新机制，但累积不对称取决于事件频率和修复机制可用性（Baumeister 等人，2001年；Hopko 等人，2022年）。

P3：信任校准取决于透明度机制；过度信任会造成安全风险，而信任不足则会降低效率（de Visser 等人，2020年；Lee 和 See，2004年）。

P4：性情因素会对直接经验所形成的习得信任产生调节作用，但不会取而代之（Hoff 和 Bashir，2015年；Loizaga 等人，2024年）。

2.2 三因素信任框架

Hancock 等人（2021年）对142项实证研究的荟萃分析（N = 7,458名参与者）提供了关于影响人机交互中信任的因素最全面的定量综合分析。他们的框架将信任前因划分为三类和六个子类别，每一类及子类别都对应着通过实证研究得出的效应量。表1列出了这些基准相关系数，可作为我们模型的验证目标。

表1。信任前因的元分析效应量（Hancock 等人，2021年）

类别	子类别	相关系数（r）	95%置信区间	研究（k）	样本（N）
机器人相关	性能	0.60	[0.54, 0.65]	66	3,471
机器人相关	属性	0.45	[0.37, 0.52]	24	1,277
人类相关	能力	0.27	[0.17, 0.36]	12	618
人类相关	特征	0.22	[0.14, 0.30]	19	1,032
环境	团队因素	0.27	[0.15, 0.38]	8	483
环境	任务因素	0.17	[0.07, 0.27]	13	577

荟萃分析揭示了若干重要模式。机器人性能是最强的预测因素（r = 0.60），这表明系统的实际可靠性在信任形成中起主导作用。

机器人属性——包括透明度、可预测性和拟人化特征——表现出中等效应（r = 0.45）。人为因素与信任的关系较弱，但具有显著性；其中，操作员能力（r = 0.27）以及信任倾向等个体特征（r = 0.22）都会影响信任差异。团队动态（r= 0.27）和任务特征（r = 0.17）等环境因素则构成了这一框架的最后部分。

值得注意的是，Hancock 等人（2021）记录了信任不对称性：消极体验造成的信任下降幅度，大于同等正面体验带来的信任提升幅度。这种不对称性对系统设计和失败恢复有重要影响，因为失败后重建信任，需要在较长时期内持续保持良好的性能。

2.3 机器人性能与可靠性

机器人可靠性——即系统成功完成指定任务的概率——是影响信任的主要因素。Desai 等人（2013年）证明，操作员会在大约 20 次交互中根据观察到的可靠性校准信任，但校准准确性取决于反馈清晰度。Salem 等人（2015年）发现，行为不一致对信任的损害比偶尔失败更严重（d = 1.24），这表明可预测性会调节可靠性带来的影响。近期研究探讨了可靠性与信任之间的非线性关系。Yang 等人（2017）观察到，信任不会随着可靠性的提高而线性增长；相反，操作员似乎会采用基于阈值的评估，在可靠性达到 70-80% 左右时，信任会出现明显变化。这一发现表明，超过某些阈值后，进一步提高可靠性带来的信任回报可能会递减。

2.4 机器人属性：透明度与个性

透明度——使机器人意图、推理和局限性对操作员可见——已受到大量研究关注。Chen 等人（2014年）的基于情境意识的智能体透明度（SAT）模型确定了三个透明度级别：当前状态、推理过程和结果预测。更高的透明度通常会改善信任校准，但过多的信息可能带来认知负荷成本（Lyons 等人，2017）。

机器人个性因素，包括感知温暖度和能力，会影响信任，且不受实际性能影响。与缺乏这些特征但功能相当的系统相比，会表现社交线索并承认不确定性的机器人获得更高的信任评分（Gombolay 等人，2015年）。Tower 和 Brooks（2024年）发现，主动承认故障比被动解释多保留41%的信任，凸显了沟通时机的重要性。

2.5 人为因素：能力、特征与经验

操作员专业知识会影响基线信任水平和校准准确性。专家的信任校准速度更快，对可靠性变化也更敏感，而新手的信任反应则更加多变（de Visser 等人，2020年）。信任倾向上的个体差异——一种稳定的人格特征——能够预测初始信任水平，但随着操作员积累对特定系统的直接经验，其影响会逐渐减弱。

机器人系统使用年限是影响信任动态的一个尤为重要的人为因素。机器人使用经验丰富的操作员，其信任反应更加校准：面对单次结果时，信任变化较小；但从长期来看，信任水平与实际可靠性之间的吻合度更高（Schaefer et al., 2016）。这种经验效应表明，新手操作员与经验丰富的操作员，其信任动态存在系统性差异。

2.6 环境因素：团队与任务

共享心智模型、沟通模式和集体经验等团队层面因素，会在个体层面影响之外，进一步影响信任发展。Sanders et al.（2019）发现，与临时组建的团队相比，建立了协作协议的团队在机器人故障期间表现出更稳定的信任水平。任务特征——复杂性、风险和时间压力——会调节各种信任前因的重要性；高风险任务会放大透明度的影响（Chen et al., 2014）。

2.7 人机交互中的基于智能体的建模

基于智能体的模型在研究人机团队方面具有独特优势。通过将个体表示为具有异质特征的主体，并让信任从互动中涌现，基于智能体的模型能够捕捉聚合统计模型无法呈现的动态（Lewis et al., 2018）。此前，基于智能体的模型在人机交互领域的应用研究涉及团队协调（Korsah et al., 2013）、任务分配（Claure et al., 2020）以及基于信任的决策（Xu & Dudek, 2015）。

然而，现有仿真模型存在一个关键局限：缺乏针对实证基准的系统验证。模型参数通常依据表面效度或理论一致性进行选择，而非通过校准，使其效应量与实验数据相符。这一验证缺口降低了人们对仿真预测的信心，也限制了其实践适用性。我们的方法直接针对这一局限，明确采用 Hancock 等人（2021）的元分析效应量作为验证目标。

近期的计算建模工作推动了人机团队中的信任表征。Guo 等人（2024）利用贝叶斯推断开发了 TIP（信任推断与传播）模型，用于捕捉由多个人类与多个机器人组成的团队中的直接和间接信任体验，这是针对这种配置提出的首个数学框架。Zahedi 等人（2023）通过部分可观测马尔可夫决策过程将信任纳入机器人规划，使机器人能够主动管理人类信任。Obi 等人（2024）提出了用于多机器人任务分配的期望确认信任模型，并将其与五种现有信任模型进行了比较。这些方法有一个共同局限：它们依赖分析或概率框架，并假设信任动态具有特定的函数形式。相比之下，基于智能体的建模能够让信任模式从智能体之间的异质性交互中自然涌现，而不必施加此类假设。关键在于，这些近期模型均未针对元分析基准进行系统验证。Krüger 和 Prilla（2024）特别指出：“大多数已提出的模型没有考虑人类与机器人之间的实体协作；即使考虑了，也未考虑信任动态的不对称性”（第268页）。我们的方法同时弥补这两方面的不足：通过明确的收益—损失速率差异来刻画信任不对称性，并根据 Hancock 等人（2021）报告的效应量进行验证。

2.8 研究空白与模型要求

文献综述确定了构建以实证为基础的人机团队仿真的若干要求。首先，模型必须纳入 Hancock 等人（2021）框架的三个类别：机器人因素（性能、包括个性在内的属性）、人为因素（能力、特征、任职年限）和环境因素（团队动态、任务属性）。其次，模型应再现已确立的效应量——信任前因与信任结果之间的相关性应落在已发表的置信区间内。第三，模型应捕捉信任不对称性，即消极体验引发的信任变化幅度大于正面体验。第四，时间动态应体现已有记录的模式，包括初始信任校准、任职年限效应和失败恢复。最后，实现应便于从业者和研究人员使用，并提供文档齐全的代码和复现材料。

3. 模型设计

本节介绍用于指导模型行为的概念框架和智能体规格。我们围绕三种智能体类型（人类、机器人、任务）和三个动态过程（信任演变、任务执行、沟通）展开介绍。

3.1 概念框架

该模型描述了一个在人机团队在共享工作空间中执行任务的情境。人类对机器人能力持有信任信念，这些信念会根据观察到的性能和沟通而演变。机器人以固定的可靠性运行，但透明度和沟通行为各不相同。任务持续到达，协作要求也各不相同。团队绩效则随着时间推移，从这些要素的相互作用中产生。

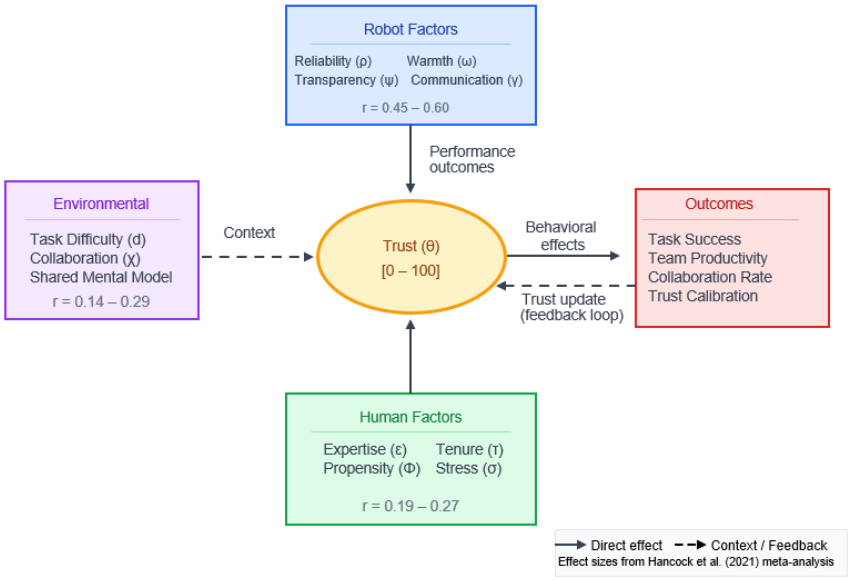


图2. 人机团队中信任动态的概念模型。

图2说明了概念关系。机器人绩效结果（成功和失败）直接影响人类信任，这种关系受透明度、人类特征（倾向性、专业知识、任职年限）和任务可见性的调节。信任进而影响人类在团队任务中与机器人协作的意愿。压力随工作负荷累积并降低人类绩效，从而在信任、协作和任务结果之间形成反馈回路。

3.2 智能体规格

表2给出了整个模型所使用的完整符号系统。我们定义了状态变量和参数，以及它们经实证论证的取值范围。

表2. 模型符号与参数规格

符号	描述	范围	来源
人类状态变量			
$\theta_i(t)$	人类 i 在时间 t 时的信任	[0, 100]	—
$\sigma_i(t)$	人类 i 的压力水平	[0, 100]	—
ε_i	人类 i 的专业知识	[20, 100]	已初始化
$\tau_i(t)$	任职时间（经验刻度）	$[0, \infty)$	累积
Φ_i	信任倾向	[0, 100]	个性
$w_i(t)$	当前工作负荷	$[0, \infty)$	计算得出
机器人状态变量			
ρ_j	机器人 j 的可靠性	[0, 100]	参数
ψ_j	机器人 j 的透明度	[0, 100]	参数
κ_j	机器人 j 的能力	[0, 100]	已初始化
ω_j	温暖感（个性）	[0, 100]	参数
γ_j	沟通频率	[0, 100]	参数
任务变量			
d_k	任务 k 的难度	[10, 100]	已生成
χ_k	协作要求	$\{0, 1\}$	生成
v_k	任务 k 的可见性	[0, 100]	生成
参数			
α	信任增益率	1.5	已校准
β	信任损失率	2.0	已校准

符号	描述	范围	来源
λ	不对称因子	1.5	Hancock 等人（2021年）
η	压力累积	0.1	Dehais 等人（2011）
δ	压力恢复	5.0	Dehais 等人（2011）

人类智能体代表具有异质性特征的团队成员。每个人类智能体都维持一个会随交互而变化的信任水平（θ）。专业知识（ε）决定任务绩效能力，其初始值取自正态分布（$\mu = 60$, $\sigma = 15$），反映技能评估中记录的劳动力异质性（Schmidt & Hunter, 1998）。信任倾向（Φ）代表个体在性格上信任他人的倾向，其初始值取自调查分布（$\mu = 50$, $\sigma = 15$），与Mayer等人（1995）的信任研究一致。任职年限（τ）随着机器人接触而累积，用于体现Schaefer 等人（2016）所记录的经验效应。

机器人智能体代表具有固定特征的自动化队友。可靠性（ρ）决定任务成功概率，并通过机器人可靠性参数进行设置。透明度（ψ）影响机器人行为和沟通对人类信任的作用方式。温暖感（ω）体现包括拟人化特征和社会临场感在内的人格因素，其依据是“计算机作为社会行动者”文献（Nass et al., 1994）。沟通频率（γ）决定机器人向附近人类提供状态更新的频率。任务智能体代表需要智能体注意力的工作项。难度（d）决定成功完成任务所需的能力。协作要求（χ）表明任务是否需要人机协作。可见度（v）决定附近人类是否能够观察到任务结果，并调节信任更新幅度。

4. 实现

4.1 平台与环境

我们使用 NetLogo 6.4.0 实现该模型。选择这一平台，是因为它已广泛用于基于智能体的社会模拟，且研究人员和从业者都能方便地使用它（Wilensky & Rand, 2015）。仿真在一个 33×33 连续环面空间中运行，智能体在其中移动、交互并执行任务。每个模拟步约代表一分钟的运行时间；一次典型运行包含 2,000 个模拟步（约相当于 33 小时的模拟运行时间）。

4.2 智能体初始化

设置过程根据用户指定的参数创建智能体群体。人类智能体被随机放置，并已初始化为：

- 信任：初始信任参数 ± 倾向性调整 专业知识：正态分布(60,
- 15)，有界 [20, 100]信任倾向：正态分布(信任倾向均值，信任
  倾向标准差) 任职年限：初始任职年限参数 压力：均匀分布(0,
- 30)
-
-

机器人智能体已初始化，参数如下：

- 可靠性：机器人可靠性参数 ± 正态分布(0, 5) 透明度：
- 机器人透明度参数 ± 正态分布(0, 5) 温暖感：机器人温
  暖感参数 ± 正态分布(0, 10) 能力：均匀分布(60, 90) 电
- 量：100
-
-

任务生成参数如下：

- 难度：正态分布(任务难度均值，20)，有界 [10, 100]协作要求：伯努利分
- 布(协作比率 + 难度调整) 可见度：均匀分布(40, 100)
-

4.3 仿真循环

每个时步按以下顺序执行：

1. **状态更新**：更新人类压力、机器人电量和任职年限累积 2. **任务搜寻**：未分配的智能体识别可用任务并向其移动 3. **任务执行**：位于任务地点的智能体贡献工作

4. **协作**：为团队任务匹配协作者 5. **沟通**：机器人以概率方式与附近的人类进行沟通 6. **信任更新**：根据观察到的结果更新信任 7. **任务完成**：处理已完成的任务、判定是否成功并释放智能体 8. **任务生成**：定期生成新任务 9. **指标计算**：计算综合性能指标

4.4 参数依据

表3记录了参数来源和校准方法。

表3。参数校准来源。

参数	数值	依据
信任增益率（α）	1.5	根据 Hancock 等人（2021年）的信任轨迹校准
信任损失率（β）	2.0	校准后不对称比率为 1.3-1.7
不对称因子（λ）	1.5	Hancock 等人（2021年）的荟萃分析发现
压力累积（η）	0.1	Dehais 等人（2011）的生理数据
压力恢复（δ）	5.0	Hancock 与 Warm（1989年）的恢复曲线
压力阈值	70	耶克斯-多德森拐点
熟悉度增长率	0.002	Schaefer 等人（2016）经验效应
通信半径	5	共享工作空间中的操作邻近性
信任更新半径	10	结果观察的可视范围

4.5 代码可用性

完整实现可在 https://github.com/rkallur2/hri_paper 获取。GitHub 仓库包含：

- Trustv6.8.nlogo：Netlogo 文件
- behaviorspace-minimal.xml：BehaviorSpace 实验文件
- trustv6-r-analysis-updated.r：R 分析脚本
- .csv 格式的结果文件

5. 验证

我们依据 Hancock 等人（2021年）的元分析基准验证了该模型，并进行了敏感性分析，以确认参数效应符合预期的关系。

5.1 针对元分析基准的实证验证

主要验证工作是将模型生成的相关系数与 Hancock 等人（2021年）报告的效应量进行比较。我们通过系统地改变各项信任前因，开展仿真实验，同时测量由此产生的信任水平。对于每个参数，我们在覆盖整个参数范围的 10 个水平上分别进行了 50 次重复实验，每个因素共获得 500 个观测值。

表4展示了验证结果。对于每个信任前因，我们报告模型生成的相关系数、95%置信区间以及验证状态（模型相关系数是否落在已发表的荟萃分析置信区间内）。

表 4. 与 Hancock 等人（2021年）元分析基准的验证。

因素	Hancock r [95% 置信区间]	模型 r	Δr	状态
机器人性能				
可靠性	0.60 [0.54, 0.65]	0.59	-0.01	✓ 已验证
机器人属性				
透明度	0.45 [0.37, 0.52]	0.38	-0.07	✓ 已验证
温暖感	0.45 [0.37、0.52]	0.34	-0.11	x 未已验证

沟通	0.45 [0.37，0.52]	0.41	-0.04	✓ 已验证
人类能力				
专业知识	0.27 [0.17，0.36]	0.14	-0.13	x 未验证
人类特征				
信任倾向	0.22 [0.14, 0.30]	0.15	-0.07	✓ 已验证
任职年限	0.22 [0.14，0.30]	0.06	-0.16	x 未验证
环境团队				
协作率	0.27 [0.15，0.38]	0.39	+0.12	x 未验证

基于主体的模型对照 Hancock 等人（2021）的元分析效应量进行了验证。8个预测变量中有4个达到了区间效度，模拟相关系数落在元分析估计值的95%置信区间内。序数效度较强（斯皮尔曼 $\rho = 0.833$），表明该模型正确再现了各预测变量效应大小的相对排序。温暖感、专业知识和任职年限的模拟效应与元分析效应之间的差异，可能反映了操作化上的不同。该荟萃分析汇总了各项研究中的多种测量，而基于主体的模型则实现了特定的行为机制。值得注意的是，协作效应超过了元分析估计值，这可能是因为仿真中的任务分配操作化比原始研究中典型的团队层面测量捕捉到了更直接的行为影响。

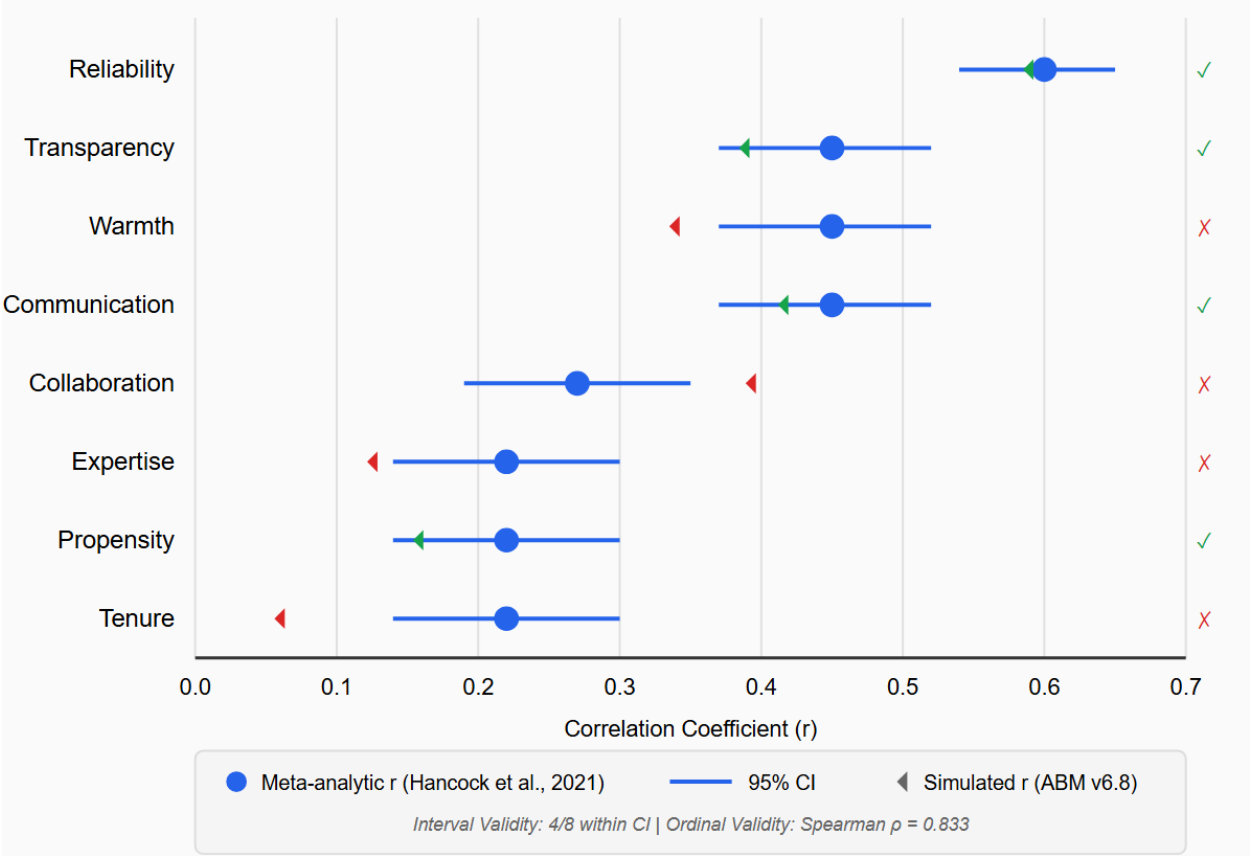


图3. For est图比较ABM模型相关性与元分析基准 .

图3展示了ABM模型相关性与Hancock 等人（2021）元分析效应量的对比。蓝色圆圈表示来自Hancock 等人（2021）的元分析效应量。蓝色水平线表示95%置信区间。三角形表示来自ABM模型v6.8（8th 版本）的模拟相关性。绿色（✓）表示模型相关系数r位于Hancock 等人（2021）的95%置信区间内，因此已验证。红色（x）表示模型相关系数r位于Hancock 等人（2021）的95%置信区间之外，因此未验证。预测变量按元分析效应量排序。底部摘要注明4/8的区间效度以及较强的序数效度（斯皮尔曼ρ = 0.833）。

5.2 信任不对称验证

一项关键的理论预测是，信任的下降速度快于其建立速度。我们通过计算所有模拟运行中累积信任损失与累积信任增益的比率来检验这一点。表5呈现了不同可靠性条件下的不对称结果。

表 5. 信任不对称验证。

可靠性	平均不对称比率	SD
30%	0.53	0.15
50%	0.37	0.11
70%	0.22	0.07
90%	0.13	0.04
总体	**0.31**	**0.18**

信任不对称性随机器人可靠性呈系统性变化。平均不对称比率从30%可靠性时的0.53（标准差 = 0.15）降至90%可靠性时的0.13（标准差 = 0.04），表明随着失败频率下降，累积信任增益日益超过损失。值得注意的是，没有任何可靠性条件产生的不对称比率接近1.50这一荟萃分析基准（Hancock 等人，2021），总体均值为0.31（标准差 = 0.18）。这一模式表明，事件频率调节了单次事件不对称性的累积表现。尽管负面事件被设定为对信任的影响是正面事件的1.5倍，但在较高可靠性水平下失败相对稀少，使信任形成占据主导。该发现表明，元分析估计值中观察到的不对称效应，可能反映的是失败率更高、或修复信任的机会少于此处所仿真情境的背景。

5.3 敏感性分析

我们采用单因素逐次分析（OFAT）和全因子设计进行敏感性分析，以刻画参数对模型输出的影响。

单因素逐次分析（OFAT）

对于八个主要参数中的每一个，我们在其他参数保持默认值的情况下，在其完整取值范围内进行变化。表6列出了各参数对关键结果变量的效应量（η2）。

表6。 OFAT 敏感性分析结果

参数	信任（η2）	任务成功（η2）	生产力（η2）
机器人可靠性	0.35	0.93	0.89
机器人透明度	0.17	0.03	0.03
机器人温暖感	0.13	0.02	0.01
沟通频率	0.20	0.01	0.00
人类专业知识	0.02	0.32	0.71
信任倾向	0.04	0.01	0.01
协作 rate	0.17	0.01	0.01
初始任期	0.01	0.03	0.02

OFAT 敏感性分析揭示了不同结果变量各自不同的驱动因素。信任形成受到多种机器人相关因素的影响，其中可靠性（η2 = .35）、沟通（η2 =.20）、透明度（η2 = .17）和协作（η2 = .17）均产生了显著的方差贡献。相比之下，任务成功和生产力主要受可靠性（分别为 η2 = .93 和 .89）以及人类专业知识（分别为 η2 = .32 和 .71）影响。社会因素对绩效结果的直接影响极小（η2 ≤ .03），尽管其对信任有显著影响，这表明信任可能发挥中介机制的作用。未来研究有必要通过正式的中介分析检验这一假设。

全因子分析

我们开展了一个34全因子实验，将可靠性（40%、70%、90%）、透明度（30%、60%、90%）、沟通频率（20%、50%、80%）和协作率（20%、50%、80%）进行交叉组合，每个单元格重复实验30次（n = 2,430次运行）。方差分析结果见表7。

表7. 信任结果的析因方差分析结果。

效应	df	SS	MS	F	p	η^2
主效应						
可靠性（R）	2	14418.82	7209.41	382.30	< .001	.171
透明度（T）	2	6729.24	3364.62	178.42	< .001	.080
沟通 (C)	2	11081.62	5540.81	293.82	< .001	.131
协作 (L)	2	6107.22	3053.61	161.93	< .001	.072
双向交互						
R × T	4	212.01	53.00	2.81	.024	.003
R × C	4	115.07	28.77	1.53	.192	.001
T × C	4	84.59	21.15	1.12	.345	.001
R × L	4	21.97	5.49	0.29	.884	< .001
T × L	4	93.64	23.41	1.24	.291	.001
C × L	4	281.27	70.32	3.73	.005	.003
三重交互						
R × T × C	8	336.80	42.10	2.23	.023	.004
R × T × L	8	100.49	12.56	0.67	.722	.001
R × C × L	8	173.71	21.71	1.15	.325	.002
T × C × L	8	186.11	23.26	1.23	.275	.002
四重交互						
R × T × C × L	16	317.36	19.83	1.05	.397	.004

效应	df	SS	MS	F	p	η^2
残差	2349	44297.79	18.86			.524

注。 N = 2,430 次模拟运行。R = 机器人可靠性，T = 机器人透明度，C =沟通频率，L = 协作率。每个因素均设有三个水平。η2 =η平方（所解释的总方差比例）。

效应汇总

效应类型	$\Sigma\ \eta^2$	解释
主效应	.454	方差的 45.4%
双向交互	.010	方差的 1.0%
三重交互作用	.009	方差的 0.9%
四重交互	.004	方差的 0.4%
残差	.524	52.4% 未解释

四个主效应均达到显著水平（$p < .001$），其中可靠性对信任的影响最强（$\eta2 = .171$），其次是沟通（$\eta2 = .131$）、透明度（$\eta2 = .080$）和协作（$\eta2 = .072$）。交互效应很小，仅有三项达到显著水平：R × T（$p = .024$）、C × L（$p = .005$）以及 R × T × C（$p = .023$）。较小的交互作用效应量（$\eta2 < .005$）表明，参数对信任的影响主要是加和性的，而非协同性的。

析因方差分析通过主效应和交互作用解释了47.6%的信任方差，其余方差则归因于基于智能体的仿真所固有的随机动态（例如，概率性任务结果、信任漂移）。这种划分与模型的设计一致：模型引入随机性，以捕捉人机信任形成过程中的真实变异。

6. 仿真场景与结果

本节考察了人机团队在五种情景下的团队绩效，这些情景代表了人机团队在部署中面临的常见挑战。我们首先定义所有情景（第6节.1），然后呈现结果（第6节.2），并对各情景的研究发现进行综合分析（第6节.3）。

6.1 场景定义

所有场景均使用 5 名人类智能体和 5 个机器人智能体，持续 2000 个时步。我们对每个场景进行了 50 次重复实验，以确保获得稳定估计。表 8 展示了完整的场景规格。

表 8. 场景参数规格。

参数	**基线**	**信任恢复**	**High 工作负荷**	**不可靠**	**最优**
初始任务	30	30	60	30	30
机器人可靠性	70%	90%	70%	40%	82%
机器人透明度 50%		80%	50%	30%	75%
机器人自主性	70%	70%	30%	90%	70%
沟通频率	40%	70%	40%	40%	50%
协作率	40%	40%	80%	40%	50%
初始信任	50	20	50	70	50
初始任期	0	0	0	0	500

情景设定依据：

基线 以适中的参数值作为参考性能，代表未进行特定优化的典型部署。

信任恢复 检验团队在配备高可靠性机器人（$\rho = 90\%$）并增强透明度（$\psi = 80\%$）时，能否从最初较低的信任水平（$\theta 0 = 20$）中恢复。该情景探讨向具有先前的负面经历的团队部署机器人这一实际问题。

高工作负荷 考察在诱发压力的条件下的性能，此时任务负荷加倍（60项任务）、协作要求较高（80%），且机器人自主性降低（30%），因而需要更多人工监督。该情景检验压力下性能退化阈值。

不可靠自动化 探讨了在机器人可靠性较差（$\rho = 40\%$）、透明度较低（$\psi = 30\%$）且自主性较高（90%）时的信任动态。该最坏情形用于检验是否会发生信任崩溃，以及系统能否恢复。

最优 测试了一种根据先前的敏感性分析设计、旨在最大化性能的配置：可靠性为82%（高于收益递减阈值）、透明度为75%（足以建立信任）、中等程度的协作（50%），以及经验丰富的操作员（$\tau = 500$）。

6.2 情景结果

表9总结了全部五种情景下的指标。

表9. 跨情景性能比较。

指标	基线	信任恢复	高工作负荷	不可靠机器人	最优
任务成功率	50.76	63.78	50.86	33.41	56.96
信任均值（标准差）	58.62 (8.7)	38.20 (8.6)	61.36 (8.2)	73.23 (8.0)	62.60 (9.4)
生产力	3.28	4.29	3.42	2.15	3.76
利用率	99.6	99.6	99.4	100.0	99.6
校准误差	11.45	52.05	8.93	33.20	19.54
不对称比率	0.225	0.069	0.164	0.552	0.107
CPI	0.662	0.700	0.676	0.591	0.714

注：括号内为标准差。*CPI* = 综合绩效指数。

6.2.1 基线情景

平均任务成功率为 50.76%，最终信任度平均为 58.62（标准差 = 8.7）。团队生产力稳定在每 1000 个智能体时步完成 3.28 项任务。信任校准误差平均为 11.45 分，表明主观信任与实际机器人能力之间具有较合理的一致性。

信任不对称比为 0.225，表明在基线条件下，累积信任增益超过了损失。团队利用率平均为 99.6%。综合绩效指数（CPI）为 0.662，作为场景比较的基准。

6.2.2 信任恢复情景

信任恢复情景呈现出一种反直觉模式：最高的任务成功率（63.78%）和生产力（每 1000 个智能体时步完成 4.292 个任务）与最低的平均信任度（38.20，标准差为 = 8.6）同时出现。尽管信任水平比基线低 35%，生产力仍比基线提高了 31%。这一发现表明，增强的透明度和沟通机制使人机协作能够有效进行，而无需较高的信任水平——人类可以直接验证机器人行为，而不是依赖基于信任的接受。然而，该情景的校准误差在所有情景中最高（52.05），表明相对于机器人的实际性能，参与者存在显著的信任不足。不对称比率降至 0.069，为观测到的最低值，反映出在这一高透明度环境中信任衰减极小。CPI 为 0.700，高于基线，表明在具备适当支架机制的情况下，低信任并不一定会损害团队效能。

6.2.3 高工作负载场景

尽管任务需求增加，任务成功率（50.86%）仍与基线相当。平均最终信任度为 61.36（标准差 = 8.2），略高于基线。值得注意的是，该情景的校准误差最低（8.93），表明在所有条件下，信任与实际系统性能之间的一致性最高。团队生产力为每 1000 个智能体时步完成 3.424 项任务，较基线小幅提升 4.4%。团队利用率仍然较高，为 99.4%。信任不对称比为 0.164，CPI 达到 0.676。高工作负载场景表明，只要系统可靠性保持稳定，任务需求增加未必会损害信任动态；相反，持续互动可能有助于实现更准确的信任校准。

6.2.4 不可靠机器人情景

不可靠机器人情景揭示了严重的信任校准失效。尽管该情景的任务成功率最低（33.41%），生产力也最低（每 1000 个智能体时步完成 2.151 项任务），但其平均信任度最高（73.23，标准差 = 8.0）——这是一种危险的过度信任状态。校准误差达到 33.20，在所有情景中排名第二，表明尽管有客观证据显示机器人性能不佳，人类操作员仍保持较高的信任度。不对称比率升至 0.552，为观测到的最高值，反映出失败频繁时信任衰减更为明显；即使比率升高，仍低于 1.0，说明

信任形成机制在一定程度上缓冲了消极体验的累积。团队利用率为 100%，表明操作员仍继续与不可靠机器人互动，并未停止使用。CPI 为 0.591，是所有情景中最低的。这一模式揭示了人机团队中过度信任的风险：持续与不可靠自动化系统互动，可能导致虚高信任在反复失败后仍然持续，进而造成不当依赖和安全关键错误。

6.2.5 最优配置场景

最优配置场景通过平衡多个绩效维度，实现了最高的综合绩效指数（CPI = 0.714）。任务成功率为 56.96%，比基线高 12%。信任水平稳定在 62.60（标准差 = 9.4），校准误差为 19.54，高于基线，但显著低于信任恢复和不可靠机器人情景。团队生产力达到每 1000 个智能体时步完成 3.756 项任务，比基线提高 14.5%。团队利用率为 99.6%。信任不对称比为 0.107，表明信任动态稳定，增益显著高于损失。重要的是，尽管生产力较低（3.76 对 4.29），最优情景的 CPI（0.71 对 0.70）仍高于信任恢复情景，因为其信任校准效果更优。这一发现表明，优化人机团队绩效需要关注信任校准，而不能仅仅单独追求生产力或信任水平的最大化。

6.3 跨场景综合

情景分析揭示了人机信任校准中的关键动态。信任恢复情景的任务成功率（63.78%）和生产力（4.29）最高，但平均信任度最低（38.20），这表明，提升透明度并采用沟通机制，无须高信任水平，也能实现有效协作。相反，不可靠机器人情景的信任度最高（73.23），尽管其任务成功率最低（33.41%）；这代表一种潜在危险的过度信任状态，其校准误差为33.20。最优配置在平衡任务成功率（56.96%）、适度信任度（62.60）和可接受的校准误差（19.54）后，取得了最高的综合绩效指数（CPI =0.714）。这些发现强调，最大化信任并不等同于优化团队绩效；相反，恰当的信任校准——即主观信任与系统客观能力相一致——才是人机团队的一项关键设计目标。不同情景下不对称比率的大幅变化（0.069至0.552）进一步支持了边界条件假说：累积信任动态取决于情境因素，包括机器人可靠性水平以及信任修复机制的可用性（Zhang等人，2023年）。

7. 讨论

7.1 理论贡献

本研究从四个方面推进了对人机团队信任动态的理解。首先，我们证明，基于智能体的模型经实证基准校准后，可以近似再现元分析效应量。该模型的8个预测变量中有4个达到区间效度，其模拟相关性落在Hancock 等人（2021年）报告的95%置信区间内。值得注意的是，序数效度表现良好（斯皮尔曼 $\rho = 0.833$），表明模型正确再现了各预测变量效应大小的相对排序——可靠性对信任的影响最大，其次是透明度、沟通和温暖感，这与荟萃分析结果一致。这种双重验证方法——同时评估绝对效应大小和相对排序——为未来的仿真研究提供了一种方法论范式。

其次，该模型揭示了信任不对称性的一项边界条件，拓展了现有理论。与荟萃分析得出的 1.50 估值（Hancock 等人，2021）相反，各情景中出现的不对称比率介于 0.069 至 0.552 之间，表明累积信任增益始终超过损失。尽管将单次事件不对称性设为 1.5×（负面事件造成的信任影响是正面事件的 1.5 倍），仍出现了这种差异。这一发现表明，当信任形成和修复路径持续发挥作用时，单次事件不对称性并不一定会导致累积不对称。随着可靠性的变化，不对称比率呈现系统性变化：从可靠性为 30% 时的 0.53 降至可靠性为 90% 时的 0.13。这表明，事件频率会调节单次事件影响差异在累积层面的表现。因此，荟萃分析得出的不对称性估值，可能反映了失败率更高或信任修复机会少于本研究所模拟情景的测量情境。

第三，情景分析显示，信任程度与团队绩效可能存在显著差异，这对“信任越高就必然越有利于人机协作”的假设提出了挑战。尽管信任恢复情景的平均信任度最低（38.20），但其任务成功率最高（63.78%），生产力也最高（4.292）；而不可靠机器人情景的信任度最高（73.23），任务成功率却最低（33.41）。这种信任与绩效脱钩现象凸显了信任校准的重要性——即主观信任与系统客观能力相一致。信任校准是有别于信任程度的一项设计目标。校准误差是重要的诊断指标：信任不足（信任恢复情景：52.05）意味着未能实现的绩效潜力，而过度信任（不可靠机器人情景：33.20）则意味着安全风险。

第四，敏感性分析发现，不同结果对应着不同的预测因子特征，这表明信任可能发挥中介作用。机器人相关社会因素（透明度、温暖感、沟通）显著影响信任（$\eta 2 = .17$、.13、.20），但对任务成功和生产力的直接影响微乎其微（$\eta 2 \leq .03$）。相比之下，可靠性和人类专业知识直接影响绩效结果（生产力分别为$\eta 2 = .93$和.71）。这种不对称性——社会因素会塑造信任，但在单变量分析中，信任并不会直接转化为绩效——表明信任可能是一种中介机制，机器人社会特征可通过该机制间接影响团队结果。这一假设值得在未来研究中通过正式的中介分析加以检验。

7.2 系统设计的关键阈值

分析揭示了若干具有实际意义的阈值：

可靠性阈值。机器人可靠性是信任（单因素分析中为 $\eta 2 = .35$）和绩效结果（任务成功为 $\eta 2 = .93$，生产力为 .89）的主要驱动因素。在测试范围（10%–100%）内，可靠性与信任之间始终呈正向关系，且可靠性水平越高，校准误差越低。关键的是，低可靠性（30%）会导致危险的过度信任：主观信任（73.23）显著高于根据实际表现所能合理信任的程度（任务成功率为 33.41%）。

校准阈值。信任校准误差低于 12 分（在基线和高工作负荷情景中观察到）表明信任与能力之间较为吻合。校准误差超过 30 分（信任恢复：52.05；不可靠机器人：33.20）则代表令人担忧的校准失调，且方向相反——分别为信任不足和过度信任。

透明度—沟通交互作用。析因方差分析发现显著的 R × T 交互作用（$p = .024$）和 R × T × C 三重交互作用（$p = .023$），表明透明度和沟通对信任的影响取决于可靠性水平。然而，交互作用效应量较小（$\eta 2 < .005$），这表明参数对信任的影响主要是累加的，而非协同的。

信任下限。观察到的最低信任水平（信任恢复条件下为 38.20）并未损害团队绩效，这表明在具备适当支架机制（透明度、沟通）的情况下，即使信任水平较低，也能实现有效协作。这一发现对人机组队所需最低信任水平的既有假设提出了挑战。

7.3 实践建议

基于我们的研究发现，我们针对人机团队设计提出以下循证建议：

优先考虑校准，而非信任程度。 系统设计人员应以实现恰当的信任校准为目标，而非一味提高信任水平。最优情景的综合绩效指数最高（0.714），并非因为信任程度最高（62.60），而是因为在任务绩效出色的同时，校准误差处于可接受范围（19.54）。在部署期间监测校准误差，有助于识别信任不足（低效率）和过度信任（安全风险）两种情况。

加大对透明度机制的投入。 透明度对信任有显著贡献（$\eta 2 = .17$），并且如信任恢复情景所示，即使信任水平较低，也可能促成有效协作。通过解释机器人行为及其局限性的透明度机制，可以在一定程度上替代基于信任的接受，使操作员能够核实机器人行为是否恰当，而非仅凭假设。

认识到可靠性是首要杠杆。 可靠性对所有结果变量都具有主导作用（$\eta 2 = .35–.93$），证实了其对人机团队成功的核心意义。若可靠性低于可接受阈值，社会因素（透明度、温暖感、沟通）便无法弥补绩效不足。投资于可靠性改进，可同时提升信任、任务成功率和生产力。

考虑信任与绩效脱钩。培训项目应使操作员为信任与性能出现分歧的情景做好准备。操作员应理解，伴随高透明度的低信任可能优于可见度有限的高信任，并且在性能不佳时仍维持信任代表着需要干预的校准失效。

针对加性效应进行设计。观察到的最小交互效应（$\Sigma \eta 2 = .023$）表明，参数改进是以加性方式而非协同方式叠加的。因此，从业者可以独立优化各项参数，而无需复杂的交互建模，从而简化系统设计决策。

7.4 局限性

当前研究受到若干局限性的制约。首先，模型的 8 个预测变量中有 4 个达到了区间效度；温暖感、专业知识、任职年限和协作的结果则落在荟萃分析

置信区间之外。这些差异可能反映了基于主体的模型中所实现的行为机制与荟萃分析所汇总的异质性测量方法之间，在操作化方面存在差异。较强的序数效度（$\rho = 0.833$）在一定程度上缓解了这一问题，证实了效应相对排序正确。

其次，仿真的不对称比率（总体为 0.31）与元分析估计值（1.50）之间的差异需要谨慎解释。这种差异并不代表验证失败，而是通过识别调节因素拓展了不对称理论。可靠性水平与不对称比率之间存在系统性关系（可靠性为 30% 时 $r = 0.53$，降至 90% 时 $r = 0.13$），这表明累积信任动态取决于单次事件影响与事件频率之间的相互作用——这一细微之处并未体现在汇总的元分析估计值中。第三，智能体的能力是固定的；而现实中的人机团队涉及学习与适应，这些都未纳入我们的模型。信任恢复情景中看似矛盾的低信任／高性能模式，在纵向部署中可能有所不同，因为信任与能力会共同演变。

第四，验证依据的是元分析汇总结果，而非特定的实验重复。尽管我们的模型再现了相关系数大小和序数排名，但无法充分验证信任演化曲线随时间变化的形态。实验室研究通常报告离散时间点的汇总信任水平，而非连续轨迹。第五，析因方差分析中的残差方差（52.4%）既反映了预期的随机动态，也可能包含未建模因素。虽然这一比例符合基于智能体的仿真的预期，但也表明，仍有相当一部分信任方差可归因于操控参数以外的因素。

7.5 未来方向

若干扩展将提升该模型的适用性。首先，应该开展正式的中介分析，以检验信任是否在社会因素（透明度、温暖感、沟通）与绩效结果之间发挥中介作用，这一点得到了单因素分析中观察到的差异敏感性模式的支持。其次，信任不对称性的边界条件值得进行实证检验。通过操纵单次事件影响不对称性和事件频率的实验室研究，可以确定累积不对称是否取决于交互时长和修复机会的可获得性，正如我们的仿真结果所表明的那样。

第三，纳入学习机制将有助于探索长期部署过程中信任校准如何演变。在不可靠机器人情景中观察到的过度信任，可能会随着长期接触而减弱、持续存在或加剧——这类动态对安全具有重大

影响。第四，整合自然语言处理技术，可模拟对话中的透明度，从概率式状态更新拓展到更丰富的人机通信。第五，通过真实人机团队开展实地验证，将是检验模型预测的最有力方式，尤其是情景分析中观察到的信任与绩效脱钩及校准动态。

8. 结论

本文提出了一个以实证为基础的基于智能体的模型，用于模拟人机团队中的信任动态。该模型整合了 Hancock 等人（2021年）的荟萃分析框架，并显示出与已发表效应量的合理一致性：8 个预测变量中有 4 个具有区间效度，且所有预测变量均表现出较强的序数效度（斯皮尔曼 $\rho = 0.833$）。

研究得出了三项主要发现。第一，信任不对称性——即负面事件对信任产生不成比例影响的现象——取决于包括事件频率和修复机制可用性在内的情境因素。尽管单次事件影响差异为 1.5×，但在所有情景中，累积信任增益均超过损失（不对称比率：0.069–0.552），这一一致性发现确定了一项拓展现有理论的边界条件：单次事件影响差异并不能保证在持续交互情境中出现累积不对称。

第二，信任程度与团队绩效可能存在显著偏离，校准误差则成为一项关键诊断指标。信任恢复情景在信任程度最低的情况下实现了最高生产力，而不可靠机器人情景在性能最差的情况下产生了最高信任。这些发现对“更高信任能够普遍促进人机协作”的假设提出了挑战，并强调了适当的信任校准作为一项独立设计目标的重要性。

第三，敏感性分析显示，信任形成涉及多个以加和方式发挥作用的驱动因素（可靠性、透明度、沟通和温暖感），而绩效结果主要由可靠性和人类专业知识决定。尽管社会因素对信任有显著影响，但其对性能的直接影响微乎其微，这表明信任可能发挥中介机制的作用——这是一个可供未来研究检验的假设。

该模型为从业者提供了一种基于证据的工具，可在部署前探索不同的团队配置。通过识别信任与性能一致（最优情景）或出现分歧（信任恢复、不可靠机器人情景）的条件，仿真能够为部署规划和操作员培训提供参考。开源实现既支持研究复现，也便于实际应用。我们鼓励研究人员扩展该

模型以适用于特定领域，对提出的边界条件进行实证检验，并探究敏感性分析所提出的信任中介假设。

参考文献

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001年). 坏比好更强。《普通心理学评论》，*5*(4), 323–370.

Burgoon, J. K. (1993). 人际期望、期望违背与情绪沟通。《语言与社会心理学杂志》，*12*(1-2), 30-48.

Campagna, G., & Rehm, M. (2025年). 人机交互中的信任评估系统综述。《*ACM*人机交互汇刊》, *14*(2), 1-35.

Chen, J. Y. C., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014年). 基于态势感知的智能体透明度。美国陆军研究实验室技术报告, ARL-TR- 6905.

Chiou, E. K., & Lee, J. D.（2023年）。信任自动化：面向响应性与韧性的设计。人为因素，*65*(1)，137-165。

Claure, H., Chen, Y., Modi, J., Jung, M., & Nikolaidis, S.（2020年）。采用公平性约束的多臂老虎机：为人类队友分配资源。*HRI '20*：*ACM/IEEE*人机交互国际会议，299-308。

de Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A.（2020年）。迈向人机团队纵向信任校准理论。《国际社会机器人学杂志》，*12*，459-478。

Dehais, F., Causse, M., Vachon, F., & Tremblay, S.（2011年）。人类与自动化交互中的认知冲突：一项心理生理学研究。《应用工效学》，*43*(3)，588-595。

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013年). 机器人故障和反馈对实时信任的影响。*HRI '13: ACM/IEEE*人机交互国际会议，251258。

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). 对自动化的依赖中信任的作用。《国际人机交互研究杂志》，*58*(6), 697- 718。

Esterwood, C., & Robert, L. P. (2022, *2022*年 *31*日8月). 人机交互中的信任修复文献综述。载于 第届 *IEEE*机器人与人类交互通信国际会议（*ro-man*）（第 1641-1646 页）。IEEE。

Esterwood, C., & Robert Jr, L. P. (2023a). 三次机会，用完出局！：多次人机信任违背及修复对机器人可信度的影响。*Computers in Human behavior*，*142*，107658。Esterwood, C., & Robert, L. P. (2023b). 心智理论与人机信任修复。*Scientific Reports*，*13*(1)，9877。Gombolay, M. C., Gutierrez, R. A., Clarke, S. G., Sturber, G. F., & Shah, J. A. (2015). 人机混合团队中的决策权、团队效率与人类员工满意度。*Autonomous Robots, 39*(3)，293-312。Guo, Y., Yang, X. J., & Shi, C. (2024). TIP：多人人机团队中的信任推断与传播模型。*Autonomous Robots*，*48*(7)，20。Hancock, P. A., & Warm, J. S. (1989). 压力与持续注意的动态模型。*Human Factors, 31*(5)，519-537。Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). 影响人机交互中信任的因素荟萃分析。*Human Factors, 53*(5)，517-527。Hancock, P. A., Kessler, T. T., Kaplan, A. D., Brill, J. C., & Szalma, J. L. (2021). 机器人信任的演变：通过序贯与比较荟萃分析进行阐释。*Human Factors, 63*(7)，1196-1229。Hancock, P. A., Kessler, T. T., Kaplan, A. D., Stowers, K., Brill, J. C., Billings, D. R., ... & Szalma, J. L. (2023). 人类为何以及如何信任：荟萃分析与扩展模型。*Frontiers in psychology*，*14*，1081086。Hoff, K. A., & Bashir, M. (2015). 对自动化的信任：整合影响信任因素的实证证据。*Human factors*，*57*(3)，407-434。Hopko, S., Wang, J., & Mehta, R. (2022). 共享空间人机协作中的人因考量与指标：系统综述。*Frontiers in Robotics and AI*，*9*，799522。Kelley, H. H. (1973). 因果归因的过程。*American Psychologist, 28*(2)，107-128。Korsah, G. A., Stentz, A., & Dias, M. B. (2013). 多机器人任务的综合分类法

k

分配。《国际机器人研究期刊》，32(12)，1495-1512。Kox, E. S., van den Boogaard, J., Turjaka, V., & Kerstholt, J. H.（2024年）。旅程还是目的地：透明度与目标达成对人机团队信任的影响。《ACM人机交互汇刊》，14(2)，1-23。

Krüger, F., & Prilla, M.（2024年）。人类与工业机器人交互中的信任动态。《行为与信息技术》，43(2)，266-288。Lee, J. D., & See, K. A.（2004年）。对自动化的信任：为适当依赖而设计。《人为因素》，46(1)，50-80。Lewis, M., Sycara, K., & Walker, P.（2018年）。人机交互中信任的作用。载于 H。

A. Abbass、J. Scholz 和 D. J. Reid（编），可信自主性的基础（第135–159页）。Springer。Loizaga, E., Bastida, L., Sillaurren, S., Moya, A., & Toledo, N. (2024). 对人机协作中的信任进行建模与测量

g。《应用科学》，*14*(5), 1919. Lucas, G. M., Becerik-Gerber, B., & Roll, S. C. (2024). 校准工人对智能自动化系统的信任。*Patterns*，*5*(9). Lyons, J. B., Koltai, K. S., Ho, N. T., Johnson, W. B., Smith, D. E., & Shively, R. J. (2017). 在复杂自动化系统中构建信任。《人体工学设计》，*25*(1), 13-19. Malle, B. F., & Ullman, D. (2021). 人机信任的多维概念及测量。载于《人机交互中的信任》（第3–25页）。学术出版社。Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). 组织信任的综合模型。《管理学会评论》，*20*(3), 709-734. McEwen, B. S. (1998). 压力、适应与疾病：异稳态与异稳态负荷。《纽约科学院年鉴》，*840*(1), 33-44. Nass, C., Steuer, J., & Tauber, E. R. (1994). 计算机是社会行动者。*CHI '94*：*SIGCHI*计算机系统人为因素会议论文集，72-78. Nembhard, D. A., & Uzumeri, M. V. (2000). 手工操作的体验式学习与遗忘

d

认知任务。《国际工业人体工学期刊》，25(4)，315-326。Obi, I., Wang, R., Jo, W., & Min, B. C. (2024). 多人与多机器人任务分配中的信任动态建模与评估。《arXiv预印本 arXiv:2409.16009》。Parasuraman, R., & Riley, V. (1997). 人类与自动化：使用、误用、弃用与滥用。《人为因素，39》(2)，230-253。Robinette, P., Howard, A., & Wagner, A. R. (2017). 对机器人的过度信任概念化：为什么人们会信任一台曾经出故障的机器人？载于《自主性与人工智能：威胁还是救星？》（第129-155页）。Cham：施普林格国际出版。Rozin, P., & Royzman, E. B. (2001). 负面偏见、负面主导效应与传染效应。《人格与社会心理学评论》，5(4)，296-320。

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). 你会信任一台（有故障的）机器人吗？错误、任务类型和个性对人机合作与信任的影响。《HRI '15：ACM/IEEE人机交互国际会议》，141-148。Sanders, T. L., Oleson, K. E., Billings, D. R., Chen, J. Y., & Hancock, P. A. (2019). 人机信任模型：理论模型的构建。《人因与工效学学会年会论文集，55》(1)，1432-1436。Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). 影响对自动化的信任发展的因素之荟萃分析：对理解未来系统中的自主性的启示。《人为因素，58》(3)，377-400。Schäfer, A., Esterbauer, R., & Kubicek, B. (2024). 信任机器人：基于人类意向性的关系信任定义。《人文与社会科学通讯》，11(1)，1-12。Schmidt, F. L., & Hunter, J. E. (1998). 人事心理学中选拔方法的效度与效用：85年研究成果的实践与理论意义。《心理学公报，124》(2)，262-274。Slovic, P. (1993). 感知风险、信任与民主。《风险分析，13》(6)，675-682。Tower, D. M., & Brooks, T. R. (2024). 人机协作中主动承认故障。《人机交互期刊》，13(2)，1-28。Tsumura, T., & Yamada, S. (2024). 通过智能体的共情行为，在一系列任务中修复人类对智能体的信任。《计算机科学前沿，6》，1461131。Wilensky, U., & Rand, W. (2015).《基于智能体的建模导论：使用NetLogo对自然、社会和工程复杂系统进行建模》。MIT出版社。

Wischnewski, M., Krämer, N., & Müller, E.（2023年4月）。测量和理解自动化系统的信任校准：最先进技术现状与未来方向综述。载于*2023*年*CHI*计算系统人因大会论文集（第 1-16 页）。

Xu, A., & Dudek, G.（2015年）。OPTIMo：用于非对称人机协作的在线概率信任推断模型。*HRI '15*：*ACM/IEEE*人机交互国际会议，221-228。Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A.（2017年）。评估用户体验和系统透明度对对自动化的信任的影响。*HRI '17*：*ACM/IEEE*人机交互国际会议，408-416。Zahedi, Z., Sengupta, S., & Kambhampati, S.（2023年）。信任感知规划：对迭代人机交互中的信任演变进行建模。载于《2023年ACM/IEEE人机交互国际会议论文集》（第 210-219 页）。ACM。

Zhang, X., Lee, S. K., Kim, W., & Hahn, S.（2023年）。“抱歉，这是我的错”：修复人机交互中的信任。《国际人机交互研究杂志》，*175*，103031。