科学通报
评 述
《中国科学》杂志社
SCIENCE CHINA PRESS

人类与人工智能在社会互动中合作行为的异同与挑战

姚苏宸[1,2], 单骥[1,2], 胡理[1,2], 吕雪靖[1,2*]

1. 中国科学院心理研究所, 认知科学与心理健康全国重点实验室, 北京 100101
2. 中国科学院大学心理学系, 北京 100049
* 联系人, E-mail: luxj@psych.ac.cn

2025-05-12 收稿, 2025-06-30 修回, 2025-08-01 接受, 2025-08-01 网络版发表
国家自然科学基金(32441103, 32171077)和中国科学院青年创新促进会人才专项(2022084)资助

摘要 合作作为社会互动的基石, 不仅决定了个体的助人或避害抉择, 更深刻影响群体凝聚力与社会演化. 人类从进化与文化视角展现出“超社会”特质: 无论在小规模采集狩猎群体, 还是在大规模民族国家体系中, 人们均表现出高度的合作倾向. 而随着大语言模型等技术的发展, 新一代人工智能展现出卓越的语言理解、社会推理和情境判断能力, 并被赋予更复杂与多元的社会角色. 人机互动的变革使社会互动迈入全新阶段. 针对这一趋势, 本文系统回顾了合作研究的经典范式, 提炼出合作过程的三大关键环节: 策略构建、行为执行与评估学习. 基于此框架, 本文总结并对比了人际互动与人机互动在决策机制、合作模式与反馈适应等方面的主要发现与显著差异, 从而进一步提出从多方面优化人工智能设计与人机协作系统的建议, 如提升决策可解释性、引入动态反馈校准、规范数据伦理与跨文化验证等, 为人机关系发展和互动模式优化提供指导.

关键词 合作, 社会互动, 人工智能, 人机协作, 人际合作

在人类社会中, 合作是一种至关重要的社会行为, 体现了个体如何在权衡个人成本与获益的基础上, 选择帮助或避免伤害他人, 因此对群体稳定和社会发展具有深远影响. 社会互动不仅是合作的基石, 也是构建信任、维系群体关系的核心机制. 从进化与文化的双重视角来看, 尽管跨物种研究表明动物在社会互动中同样存在合作甚至亲社会行为[1], 人类却相比于其他哺乳动物表现出一种“超社会”属性[2,3], 其特征体现在无论是在小规模群体(如狩猎/采集群体)还是大规模群体中(如民族/国家群体), 人类都表现出更高的合作倾向与更频繁的合作行为[4~6], 而且这在童年中期就得以体现[5,7,8]. 更重要的是, 人类的合作行为常常超越血缘和地域限制, 展现出高度亲社会性, 许多行为无法简单用最大个人收益的理性模型来解释, 这也促使研究者开始关注合作的影响因素与动态机制[9,10].

近年来, 随着人工智能技术, 尤其是大语言模型(large language models, LLMs)的飞速发展, 新一代人工智能(artificial intelligence, AI)展现出卓越的自然语言理解和社会推理能力, 也被逐步赋予更复杂的社会角色, 越来越频繁地与人类进行互动. 因此, 我们正在经历一场深刻的社会互动范式变革: 互动的对象不再局限于人类, 而是扩展至智能代理(AI agents), 形成多样化的人机互动模式. 这种变革也引发了社会科学与AI领域关于人机关系与互动模式未来发展的深刻思考. 目前, AI在诸多任务中的表现越来越接近人类[11], 在一些社会行为研究中甚至被用于替代人类被试, 探索公平性、框架效应等问题[12~14]. 然而, AI的反应表现是否能够真正等同于人类的社会认知与行为反应仍不清楚.

在这样的背景下, 一个核心问题逐渐浮现: 随着社

引用格式: 姚苏宸, 单骥, 胡理, 等. 人类与人工智能在社会互动中合作行为的异同与挑战. 科学通报
Yao S, Shan J, Hu L, et al. Human–AI social cooperation: convergences, divergences, and challenges (in Chinese). Chin Sci Bull, doi: 10.1360/CSB-2025-0662

会互动对象的拓展, AI的“类人化”能力与其“非人”本质之间的矛盾如何重塑社会互动? 当前研究多局限于孤立地考察人际互动中人类的合作决策, 或AI在与人互动时的响应特征与学习机制, 缺乏一个系统性的理论框架对不同决策主体在合作行为中的异同及机制进行对比. 这不仅限制了相关理论的发展与深化, 也对促进人机互动的实践带来了挑战. 为更好地理解人际互动与人机互动在合作模式及交互过程中的关键异同, 本文首先对复杂社会情境下合作行为的研究范式进行梳理; 然后基于共同的博弈范式提出阶段对比框架(策略构建、行为执行、评估学习), 比较人类与AI两类不同决策主体在社会互动中合作行为的核心机制与行为模式; 最后探讨不同互动情境下人类和AI的合作决策机制, 并为人机关系与互动模式未来发展提出建议.

1 复杂社会情境下合作行为的研究范式

博弈论(game theory)作为一种数学理论, 最初用于评估个体在互动中的行为选择, 尤其是在面对不确定性和利益冲突时的决策过程[10,15]. 由于能有效抽象表达现实中的复杂情境, 博弈论迅速成为社会科学领域中常用的探索工具[16], 广泛应用于经济学、政治学、社会学和心理学等领域[17]. 在心理学中, 研究者基于博弈理论发展出了多种博弈游戏范式用于探讨个体和群体如何在互动中产生行为变化, 为社会合作的研究提供了坚实基础. 通过博弈实验范式, 研究者能够探讨个体在复杂社会情境中的决策模式并推广至诸多现实情境.

具体而言, 独裁者博弈(dictator game, DG)和囚徒困境(prisoner’s dilemma, PD)等经典范式已被广泛应用于合作行为的研究[18~21]. 在独裁者博弈中, 有一名“独裁者”负责决定如何在两个分配选项之间进行选择, 另一名“接受者”则被动接受独裁者的决定. 这一范式常用于探讨个体的公平偏好、社会责任感以及利他行为的动机. 类似地, 对于人机合作, 研究者通过将机器设定为“独裁者”或“接受者”, 检验其是否能表现出与人类相似的公平行为[15]. 囚徒困境也是研究合作或背叛决策的常用范式. 在此范式中, 两名玩家需要在不知道对方策略的情况下, 选择“合作”或“背叛”. 典型的囚徒困境一般有四种可能的结果: (P) 双方都选择背叛, 双方都获得较低回报; (R) 双方都选择合作, 双方都获得中等回报; (T) 自己选择背叛, 对方选择合作, 自己获得最高回报, 而对方没有任何回报; (S) 自己选择合作, 对方选择背叛, 自己没有任何回报, 而对方获得最高回报. 该范式认为, 理论上, 基于收益最大化原则, 背叛应当是最优选择(T>R>P>S), 因为选择背叛能够带来更高的回报, 但实际情况却并非如此. 除了用于模拟短期决策冲突, 该范式也常用于研究声誉、惩罚和奖赏机制如何影响和促进长期的持续合作[22,23]. 在人机合作研究中, 囚徒困境也常用于评估AI在面对不确定性和互动反馈时的策略选择[16,24].

在这些研究范式的基础上, 随着研究方法的不断扩展, 更多复杂的博弈范式逐渐涌现, 例如联盟的形成以及议价博弈(如最后通牒任务和信任博弈等). 这些博弈范式进一步拓展了合作研究范围, 例如如何协调在需要所有或大多数参与者达成一致的情境下所存在的不同利益和意见等[25]. 以信任博弈为例, 该范式主要用于探究信任与背叛之间的动态关系. 在这一范式中, 一名玩家(投资者)需要决定将一部分资源交给另一名玩家(受托者), 而受托者随后可以选择返还部分资源或完全自我保留. 信任博弈范式通过模拟初始信任的建立及其后续回报验证, 反映和探究个体对他人信任程度和社会规范的依赖性. 在人机合作的研究中, 信任博弈也为检验AI如何获取人类的信任并预测对方的行为方面提供了有效的测试平台.

2 人际互动与人机互动在合作行为中的异同

越来越多的证据表明, 尽管人际互动与人机互动在合作行为模式和交互过程中的底层机制存在诸多不同, 但在相同的研究范式下, 其行为表现仍展现出值得关注的相似之处. 为了系统理解并对比二者的异同, 本文基于博弈理论的核心要素: 主体属性、决策奖励以及决策结果反馈[21], 构建一个包含策略构建、行为执行以及评估学习的三个阶段对比框架(图1). 人类个体和AI系统都遵循着从策略生成、实际行动到反馈学习这一动态循环. 在这一连续框架中, 人类与AI不仅在各阶段展现出不同的认知特性与行为模式, 阶段之间的衔接方式亦体现出结构性差异. 特别是在行为执行过程中, 一些非理性因素(如情绪反应、认知偏差或动机冲突)对人类合作行为的调节作用尤为显著; 而AI系统则更多依赖于预设算法执行既定策略, 或根据反馈在评估阶段进行延迟调整. 因此, 本研究将这些动态调节因素主要纳入“行为执行”阶段加以分析, 强调其在实际交互过程中的显性影响, 而非其在前期策略生成中

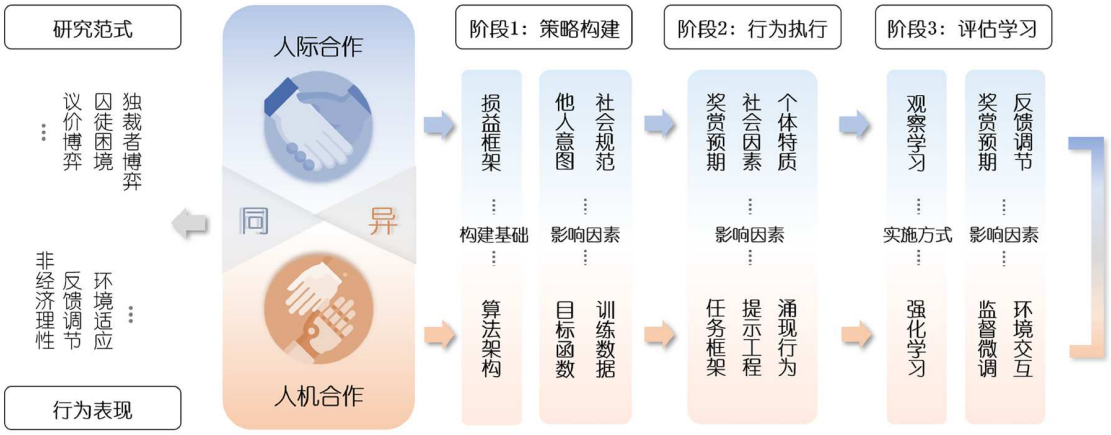

图 1 (网络版彩色)基于决策过程的人际互动与人机互动中合作行为的对比框架(图片改自网站https://www.flaticon.com/)

Figure 1 (Color online) A comparative framework of cooperation based on decision processes in human-human and human-AI interactions (Elements sourced from https://www.flaticon.com/)

的潜在介入，从而系统对比人际合作与人机合作在行为模式和交互过程中的关键差异.

2.1 策略构建阶段

在人际合作中，个体的合作动机既受到人类进化过程中逐步形成的基本心理机制的影响，也受现代情境下理性策略思考的调节. 研究表明, 人类在漫长的进化过程中逐渐形成了一系列有利于群体生存的心理倾向[26]. 这些机制包括对互惠行为和群体共享利益的天然偏好[27]，帮助个体在远古社会中形成互助合作的群体. 这种生物学基础表明, 合作行为并不是现代社会的产物，而是在远古时期就已经存在的适应性行为[24,28~30]. 根据互惠利他主义的理论[30], 个体在预期未来回报的情况下, 会付出代价帮助他人, 这一机制在多项行为实验和计算建模研究中得到了验证. 例如, Fehr和Fischbacher[24]的研究表明, 在公共物品博弈中, 个体往往会牺牲部分个人利益以维护群体公平，进而促进合作; 当基于直觉决策时, 人们更倾向于依赖社会启发式, 而这种启发式是基于以往的合作经验形成的, 进而产生合作或利他意愿[31]; 与此同时，Tomasello[29]提出的共享意向性理论强调，人类具备形成共同目标和意图的独特能力，表现为个体能在合作过程中有效协调各自行为，而共享意向性有助于合作行为(如间接互惠机制)的稳固演化.

在现代社会中，理性决策过程在合作选择中十分重要. 个体在面临复杂、多变的决策情境时, 会综合考虑行为的直接成本与潜在收益，并综合逻辑推理和对未来情境的分析做出“最优”选择. 这样的决策过程不仅依赖于内在的奖赏激励和风险评估，也受到外部环境和社会规范的影响. 个体不仅可以通过直接沟通或信息传递设定共同目标，还能通过声誉机制和非正式规则(如从众偏差)对行为进行协调, 这种灵活沟通与规划的能力使得人类能够根据情境、成本–收益框架和对未来奖励的预测不断调整自身的合作策略[32~34]. 例如，个体会更倾向于与具有良好声誉或能够给他们带来好处的人建立关系[35,36]，并通过不公平分配或“以牙还牙”的策略促使潜在合作伙伴参与合作[37]. 除此之外, 在历史与文化的作用下，社会规范被内化为特定情境下的合作动机，在个体面对公共物品博弈等情境时表现为一种认为“他人会回报，我也愿意贡献”的预期和合作信念[38].

相比之下，AI的合作意图与策略选择则主要依赖于外部设计，比如预先设定的算法架构、目标函数与训练数据，而非源自内在情感或长期社会经验. 近年来, 深度强化学习(deep reinforcement learning, DRL)逐渐成为构建AI合作策略的核心方法：基于“奖励”信号最大化的目标，智能体通过与环境反复交互从而不断更新其行为策略. 进行策略构建时常用的算法包括深度Q网络(deep Q-network, DQN)、策略梯度(vanilla policy gradient, VPG)及近端策略优化(proximal policy optimization, PPO)等[39~41]. DQN利用神经网络(一种模拟人脑神经元互联结构的计算模型)拟合状态—动作

价值函数(记作Q(s,a), 即在给定状态s下采取动作a所能获得的回报), 从而在离散动作选择中实现最大化预期累积奖励[40]; VPG则直接对策略函数π(a|s)(即在状态s下采取动作a的概率分布)进行参数化, 通过梯度优化直接提升策略期望回报[42]; PPO则通过在每次策略更新时引入概率比裁剪(clipping)限制每次更新的幅度, 以防止策略更新幅度过大导致的性能崩溃, 兼顾了对新策略的探索与对旧策略的稳定保持, 在复杂任务中尤为高效[41]. 由于现实中社会互动往往涉及多个交互主体, 且每个主体的决策不仅会影响自身收益, 还可能会协同影响其他主体, 从而表现出一种动态性. 因此随着研究的不断深入, 多智能体系统(multi-agent systems, MAS)逐渐成为模拟复杂社会互动和分布式决策的重要工具. 研究者也基于此开发了如多智能体深度确定性策略梯度(multi-agent deep deterministic policy gradient, MADDPG)等分布式的决策框架与算法, 利用集中训练与分布式执行(centralized training with decentralized execution, CTDE)的范式来协调个体目标与群体收益, 并在自动驾驶协同、资源分配和协商谈判等实际场景中取得了显著成果[14,43~45].

然而, 这些基于强化学习算法的策略构建仍主要聚焦于即时收益的最大化, 缺乏模拟人类社会交互中通过信任、声誉和文化内化实现长期合作的能力. 同时, MAS依赖于Transformer注意力机制、提示工程与语境信息, 通过“链式思维”来提升推理能力, 这类系统需要结合预设角色设定和任务描述, 利用大量预训练数据和适当的温度参数调试, 从而生成决策策略并实现行为的交互[46~49]. 因此尽管MAS在谈判和让步等情境中展现出一定的社交启发式, 其本质仍是对大量数据中的决策模式进行组合, 缺乏对社会规范和内在情感动机的深层理解. 这种差异导致人类和AI在迭代囚徒困境(iterated prisoner's dilemma)中的表现差异尤为显著: 人类参与者通常采用“以牙还牙”策略, 通过条件性宽恕在遭遇背叛后修复关系, 进而维持长期合作收益; 然而, 基于Q值更新的AI智能体由于受马尔可夫决策过程短期奖励的限制, 往往更倾向于选择背叛(理论奖励最大化)策略, 即在重复互动的情景下表现出对即时收益的系统性偏好[50,51]. 此外, 传统的MAS强化学习方法在开放动态的环境中, 由于无法内化诸如利他、惩罚或声誉机制等文化规范, 容易陷入“搭便车”(即个体在不承担相应成本的前提下, 利用他人合作所带来的集体收益)或“过度竞争”的困境并走向次优均衡的局面, 而人类群体则通过长期社会学习和文化传递, 形成了较为稳定的合作模式[44,52].

2.2 行为执行阶段

在人际合作过程中, 个体的行为常常受到情感、认知偏差、计算能力限制以及有限理性等因素影响, 从而表现出复杂性和变异性[10,52~55]. 理论上, 在博弈游戏中完全理性的决策者应选择最大化自身利益, 即保留全部收益. 然而, 实验结果表明个体在决策时往往并不完全遵循纯粹的经济理性[56,57]. 大量实验证据表明, 人类在实际决策中往往倾向于合作. 例如, Mengel[58]对囚徒困境任务的荟萃分析发现, 约37%的试次里, 受试者选择合作, 而且个体在决策过程中不仅关注短期利益最大化, 还会考虑长期合作关系的稳定性及公平性、社会规范、文化背景以及群体归属等因素的多方面影响. 已有研究发现公平性与群体目标和长期合作密切相关, 不公平的出现往往会阻碍合作[59~61]. 除此之外, 个体决策受到认知、社会期望与文化背景的共同影响: 内化社会规范的个体更倾向于合作行为, 并通过惩罚机制维护群体规范[62,63]; 类似地, 引入第三方裁定机制有助于提升规范的执行力, 从而显著扩大直接互惠机制的适用范围[6,64]. 与此同时, 文化价值观显著影响个体在合作任务中的策略倾向. Herrmann等人[65]在一项跨文化研究中发现, 社会规范执行较弱的文化更易出现对合作者的“反社会惩罚”行为, 进而抑制合作的形成, 而规范体系健全的文化中则更倾向于通过惩罚机制鼓励合作并限制“搭便车”行为.

相比之下, AI的行为执行主要依赖于预先编程的算法和输入信息, 理论上不应具备人类意义上的情感与动机. 然而在MAS中, 未曾明确编程的“涌现行为”有时也会出现(即在互动中自发形成的复杂行为). 这种具有“非界定性”的行为形式, 虽然不具备人类意义上的意图与理解, 但在一定程度上增强了AI系统适应环境变化的能力[66,67]. 换言之, 尽管驱动机制截然不同, 但目前的AI, 特别是LLMs, 在行为执行层面也常不选择经济理性最优解, 而是表现出与人类相似甚至超越人类的合作与公平倾向. 例如Brookins和DeBacker发现, LLMs在长期合作框架下更倾向于合作, 而在一次性博弈中更易选择背叛, 这种条件性互惠模式与人类策略相似[68]. LLMs在独裁者博弈中甚至表现出更强的公平倾向: 人类平均保留约70%的资源, 而LLMs在70%试次中选择平分资源, 且从未出现完全自利的行为[68,69]. 在

一次性囚徒困境任务中，研究者发现LLM的合作率亦高于人类参与者[58,70]. 这些结果均表明，LLMs可能在合作行为中表现出更强的利他倾向，表现出与“非经济理性”人类相似的反应.

不过，这种合作倾向并非一致稳定：Lorè和Heydari[71]采用四种不同的一次性经济博弈对GPT-3.5和GPT-4进行了对比，发现不同版本的LLMs在合作任务中的行为模式存在显著差异，且受到任务情境的影响. 具体而言，GPT-3.5对情境变化(如“友谊”“团队合作”)更为敏感，而GPT-4则更关注游戏的逻辑推理，对情境变化的适应性较低，可能导致其合作水平在某些情境下趋向极端(即最高或最低分配策略).

由于具有较强的推理能力，LLMs可以通过“上下文少样本学习”(contextual few-shot learning)通过少量示例输入, 推断并生成符合预期的行为模式[49], 这表明提示语对LLMs的合作行为塑造起到了重要的作用. Horton[72]通过设置不同人格特质提示(如“公平偏好”“效率优先”“自利性”)，发现具有“自利性”提示的LLMs倾向于最大化自身收益，而具有“公平偏好”提示的则展现出较高的社会公平性. 类似地，Phelps和Russell[73]发现，LLMs在提示引导下能在博弈游戏中呈现利他、合作、竞争、自私等不同行为模式特征，尽管一些特征未能在模拟合作的博弈游戏中得到预期的分离. 由此可见, 当LLMs收到多样化的提示(prompt)并在一定程度上模拟人类特征时，其分配的公平性具有大幅度的变异.

除此之外, 有研究通过精确的指导语设定, 探讨了GPT-4是否能胜任在两种社会博弈任务中担任其中一方角色. 结果发现, 提示中对社会偏好的提示能显著改变GPT在游戏中的行为，使其符合具有社会偏好者的典型行为特征，且表现出与人类相似的合作发展轨迹[74].尽管如此，GPT与人类在多轮次博弈中的选择变化上仍存在差异[75]. 当要求模拟“公平的仲裁者”时，LLMs倾向于机械地平分资源，而人类仲裁者会根据贡献度和各自对资源的主观价值差异化分配，这反映出LLMs对价值理解的缺乏[76]. 这种“角色认知偏差”不仅受限于训练语料中相关概念的共现频率，还受到非语义提示属性(如标签顺序和语境框架)的影响, 进一步表明了模型在角色生成和行为执行中与人类内在机制的根本差异. 因此，尽管LLMs能够在短期内模拟出符合提示语境的合作行为，但其行为缺乏基于生物神经和社会文化的深层次调控，显示出抽象符号系统与具身经验之间的桥接问题.

2.3 评估学习阶段

在评估学习阶段，人类会根据观察到的互动结果和他人行为不断更新自身的信念，学习特定情境下的社会规范, 并推断他人的特征和意图, 从而在后续互动中调整和优化策略, 形成一种观察学习的过程[77~81]. 近年来, 计算模型被应用于探索这一动态过程. 相比于传统分析关注外显行为结果，计算模型通过拆分的方式, 能够定量描述在动态决策中难以直接测量的认知过程, 从而揭示潜在的心理计算机制[82]. 其中, 强化学习是常用的方法之一，它描绘了个体在与外界环境交互的过程中, 根据反馈调整策略与行为的学习过程. 基于行为主义理论，个体通过反复试误将刺激或行为与结果关联形成期望，并在得知结果时根据预期误差调整行为. 这一过程中，个体通过不断整合预期误差和奖励期望完成学习[83]. 行为的调整更新往往基于多维度反馈, 在合作游戏中，人们往往根据游戏结果更新对他人或自身能力的判断，并预测合作可获得的奖赏[80]. 除此之外, 由于互动往往包含多次往来, 个体在此过程中不仅会考虑当前行为的即时结果，也会考虑行为对未来互动的影响. 例如，在公共物品博弈任务中，个体会根据其他成员的行为与预期误差推断与其在下一轮中合作的可能性和即时收益，同时也会计算剩余游戏回合中的长期收益，并共同参考构成行为决策模式[47]. 最后，个体还会根据环境的不确定程度适应性地调整学习率[84].

现代人工智能系统通过数据驱动的强化学习与监督式微调机制展现出了显著的模型迭代能力. 在马尔可夫决策过程的框架下，这类系统基于与环境交互获得的奖励信号以及人类标注的监督数据持续优化自身的策略函数，从而提升在复杂决策任务中的表现效率[40]. 有研究表明，DQN等算法在Atari游戏环境(用于评估强化学习算法在多维度视觉输入下的决策能力的测试平台)中可实现每秒数百万次参数更新[85]，这种高频次优化的能力在时间尺度上显著超越了人类的学习过程. 然而, 当前在情境化认知方面仍存在理论性局限: Sutton和Barto[42]的强化学习理论框架指出，现有算法的环境建模主要依赖于可观测的特征空间，难以有效编码文化语境中的隐含维度，如社会习俗、道德规范等, 因此导致决策质量不高; Hadfield-Menell等人[86]采用逆向奖励推断研究进一步揭示了当任务涉及多维社

会规范交叉作用时，AI系统对内隐社会契约(如合作中未明文规定的“弱势群体优先”原则)的识别准确率显著低于人类专家, 量化了AI与人类的文化认知差距.

除此之外，当前基于深度学习的情感计算系统在处理动态社交线索时面临多重技术瓶颈. 现有模型在解析微表情和语调韵律等连续线索时存在显著偏差, 主要由于现有方法采用固定帧采样，导致对自发微表情快速小幅度相位变化的捕捉不佳[87]. 尽管Transformer的自注意力机制可捕获长程依赖, 具有较强的跨时空搜索能力[48], 但其二次计算复杂度高, 实际应用中通常采用分段处理方式, 这破坏了信息的连续性表征. 在跨模态情感一致性上，现有方法也仍存在架构性缺陷, 例如特征对齐偏差和动态权重僵化问题，这使得其在冲突情感识别任务中(例如“微笑但语气愤怒”)表现不佳[88].

尽管人类和AI系统在评估学习的机制上存在显著差异，但两者都具备根据互动结果和反馈信号调整自身策略和行为模式的能力. 人类通过整合预期误差和奖励期望进行学习, 并考虑长期收益; AI则通过优化策略函数最大化奖励信号，或利用人类标注数据进行微调. 这种基于反馈进行迭代优化的能力, 是实现长期人机合作的基础.

3 启示与展望

LLMs作为AI的重要分支, 由于展现出卓越的语言理解和社会推理能力，其在社会交互任务中被逐步赋予越来越重要的角色. 目前, 尽管其与人际互动中合作行为的发生机制和交互过程仍存在显著差异，但大量研究通过博弈游戏范式证明LMMs呈现出“拟人化”的趋势，即它们往往不会选择绝对理性的最优自利结果, 而是表现出较高的合作意愿. 如上文所述, 一些研究表明LMMs在这些典型研究范式中的合作概率甚至超过了人类，且其行为可以根据情境和提示工程表现出一定的变异性. 这一趋势不仅反映了技术在自然语言处理和决策支持方面的不断成熟，也预示着AI在社会互动中正逐步承担更多复杂角色. 这种趋同现象既为人机协作带来新机遇，引发深层的伦理挑战与技术哲学反思, 也对传统的合作决策理论提出了挑战: 是否能够基于已有的人际合作理论，有效解释或预测人机合作中的行为模式? 人类的合作行为植根于进化驱动、情感机制与社会文化建构，而LLMs的行为则由模型结构、训练语料与提示输入共同驱动，因此在面对行为上“拟人化”的AI系统时，人类可能会套用人际合作框架进行判断和决策. 然而，由于AI本质上的“非人性”, 一些合作理论(如基于血缘的亲缘选择理论)在此背景下可能失去适用性; 相比之下, 若AI具备足够的本土化或文化嵌入能力，强调文化规范、社会互惠或互惠利他等机制的理论仍可能在一定程度上适用. 而从AI本身来看, 其合作行为并非出于生物性动机, 而是由提示词、预训练语料与模型结构等“规则”共同驱动，因此相比于进化生物学背景下的合作理论，强调公平性、规范一致性等“超越情境的动机”所构建的合作理论可能在捕捉和预测AI在合作任务中的行为特征与变异模式体现出独特优势.

目前的研究大多聚焦于AI在合作任务中表现. 然而, 当AI成为合作对象时, 它们如何影响人类的合作行为以及人类与智能体互动中的动态协商机制仍存在认知盲区，尤其是在行为的发生机制、动态变化以及影响因素的探讨上仍显不足. 例如，对于如何理解其“黑箱”决策过程、建立信任以及更有效地开展合作等问题尚缺乏系统性阐释. 有研究表明, 与全人类构成的团队相比, 引入AI的团队中人类的整体表现可能下降, 原因可能在于AI的加入削弱了团队成员间的协调效率与信任感[89]. 此外, Traeger等人[90]发现, 与设定为中性或沉默的机器人相比，与表现出“脆弱特征”(如更多的自我暴露)的机器人交互可以显著改善后续人类团队成员之间的沟通行为. 这种非对称性揭示出人机协作中“认知对齐陷阱”: 当前基于表面行为模仿的训练范式可能导致人类对AI行为产生过度的心智化理解，将其反应错误地解释为人类情感或意图的表达[91]. 除此之外, 大多数AI的训练数据源于人类生成的语料, 因此其输出往往也不可避免地继承了人类认知偏见和错误[92,93]. 不过，相对于人类群体中的偏见，AI内部的偏见可以使用提示工程部分消除，但消除策略在很大程度上取决于研究人员识别偏见并设计干预策略的能力[94].

随着AI日益深入各个社会领域，其与人类协同工作的情境不断增多，探索新的社会交互模式已具有重大理论和实践意义，心理学将在解析这一新模式中发挥关键作用. 基于上述挑战, 未来研究可进一步关注如何建立计算模型来模拟人—AI混合团队的动态合作过程, 用于预测冲突或协同的产生. 同时, 可以进一步结合行为范式与神经影像，系统研究和比较人类在与AI或人类合作时的不同神经响应、认知加工与情绪调节

过程. 在此基础上, 可以通过设定不同的AI行为特征(如表达不确定性、展示学习过程或展现情绪反馈), 来调节人类的信任建立与合作倾向, 从而推动人机交互从“工具型”向“伙伴型”的发展路径转变.

破解这些困局, 需要技术和伦理两方面共同推进. 首先, 提高AI决策过程的可解释性至关重要. AI决策过程的“黑箱”特性(可解释性低)导致用户难以理解其决策逻辑, 进而可能导致用户走向信任危机和过度心智化两种困境. 在理论层面上, 认知对齐评估框架能基于心理学角度帮助制定“解释是否有效”的标准; 在技术层面上, 通过可视化重要特征和决策路径追踪等方法可以增强决策透明度. 其次, 系统应建立动态校准机制, 引入实时反馈, 让人类与AI在长期协作过程中能够不断调整与匹配决策策略与结果预期[95]. 例如, 在策略构建阶段, AI可以使用基于人类决策大数据训练的意图识别与预测模型, 更准确判断个体意图和偏好, 实现“因人而异”的预对齐. 协作目标与策略透明化也有助于帮助人类清晰理解AI的任务边界与可能行为区间, 从而降低心智化误解风险. 但这类透明化手段也可能反向影响人类行为, 其适用范围仍需进一步研究. 在行为执行阶段, 可以引入基于语气、表情、反应等多模态信号的自适应调节机制, 提高互动流畅性, 促进信任建立. 在评估与学习阶段, 应推动从简单的奖励机制过渡到更复杂的社会归因机制, 使AI能够在反复交互中逐步更新对他人行为的模型构建和自身的交互策略, 避免因策略突变导致合作中断或失配.

为确保上述机制符合伦理, 相关人员还需建立全过程的数据伦理框架, 从训练数据、建模逻辑到输出后果的全过程对算法偏见与社会不平等进行动态监测与回应. 同时应构建多层级隐私保护结构, 最大程度地减少个人信息泄露风险, 保障数据使用的合法性、透明性与用户知情权. 协作系统的伦理设计还应确保情感与决策边界清晰, 在高风险场景中(如医疗、自动驾驶), 应保证AI具备高可解释性、透明操作和人工接管机制; 而在伦理敏感领域(如司法、精神评估), AI的使用需受到严格限制与审查.

总之, 随着AI深入社会各领域, 人机协作面临前所未有的发展机遇与挑战.只有在技术、伦理与系统设计上同步推进, 建立透明、动态、可信的协作机制, 才能确保AI在推动社会进步的同时, 赢得大众足够的信任和理解, 为未来构建更加公正、协同和可持续的交互模式奠定坚实基础.

参考文献

1 Zhang F R, Liu J, Wen J, et al. Distinct oxytocin signaling pathways synergistically mediate rescue-like behavior in mice. Proc Natl Acad Sci USA, 2025, 122: e2423374122
2 Gowdy J, Krall L. The economic origins of ultrasociality. Behav Brain Sci, 2016, 39: e92
3 Turchin P. The puzzle of human ultrasociality: how did large-scale complex societies evolve? In: Richerson P J, Christiansen M H, eds. Cultural Evolution: Society, Technology, Language, and Religion. Cambridge: The MIT Press, 2013
4 Handley C, Mathew S. Human large-scale cooperation as a product of competition between cultural groups. Nat Commun, 2020, 11: 702
5 Jaeggi A V, Gurven M. Reciprocity explains food sharing in humans and other primates independent of kin selection and tolerated scrounging: a phylogenetic meta-analysis. Proc R Soc B, 2013, 280: 20131615
6 Mathew S, Boyd R, Van Veelen M. Human cooperation among kin and close associates may require enforcement of norms by third parties. In: Richerson P J, Christiansen M H, eds. Cultural Evolution: Society, Technology, Language, and Religion. Cambridge: The MIT Press, 2013
7 Silk J B, Brosnan S F, Vonk J, et al. Chimpanzees are indifferent to the welfare of unrelated group members. Nature, 2005, 437: 1357–1359
8 Warneken F, Chen F, Tomasello M. Cooperative activities in young children and chimpanzees. Child Dev, 2006, 77: 640–663
9 Bai Q, Chen S, Luo S. Empathy-driven group intergenerational decision-making (in Chinese). Chin Sci Bull, 2025, 70: 1079–1090 [白麒钰,陈尚仪,罗思阳. 共情驱动的群体代际决策. 科学通报, 2025, 70: 1079–1090]
10 Camerer C F. Behavioral Game Theory: Experiments in Strategic Interaction. Princeton: Princeton University Press, 2003
11 Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. arXiv, 2022: 2203.02155
12 Aher G, Arriaga R I, Tauman Kalai A. Using large language models to simulate multiple humans and replicate human subject studies. arXiv, 2022: 2208.10264
13 Bail C A. Can Generative AI improve social science? Proc Natl Acad Sci USA, 2024, 121: e2314021121
14 Zhang R, Hou J, Walter F, et al. Multi-agent reinforcement learning for autonomous driving: a survey. arXiv, 2024: 2408.09675
15 Nisan N, Roughgarden T, Tardos E, et al. Algorithmic Game Theory. Cambridge: Cambridge University Press, 2007
16 Charness G, Rabin M. Understanding social preferences with simple tests. Q J Economics, 2002, 117: 817–869

17 Samuelson L. Game theory in economics and beyond. J Econ Perspect, 2016, 30: 107–130
18 Buchholz W, Eichenseer M. Prisoner's dilemma. In: Marciano A, Ramello G B, eds. Encyclopedia of Law and Economics. New York: Springer, 2016. 1–5
19 Charness G, Gneezy U. What's in a name? Anonymity and social distance in dictator and ultimatum games. J Econom Behav Organ, 2008, 68: 29–35
20 Guala F, Mittone L. Paradigmatic experiments: the dictator game. J Socio-Economics, 2010, 39: 578–584
21 Pruitt D G, Kimmel M J. Twenty years of experimental gaming: critique, synthesis, and suggestions for the future. Annu Rev Psychol, 1977, 28: 363–392
22 Gross J, De Dreu C K W. The rise and fall of cooperation through reputation and group polarization. Nat Commun, 2019, 10: 776
23 Wu J, Luan S, Raihani N. Reward, punishment, and prosocial behavior: recent developments and implications. Curr Opin Psychol, 2022, 44: 117–123
24 Fehr E, Fischbacher U. The nature of human altruism. Nature, 2003, 425: 785–791
25 van Dijk E, De Dreu C K W. Experimental games and social decision making. Annu Rev Psychol, 2021, 72: 415–438
26 Yang Y, Tang Y, Peng W W, et al. Empathy: the genetics-environment-endocrine-brain mechanism (in Chinese). Chin Sci Bull, 2017, 62: 3729–3742 [杨业, 汤艺, 彭微微, 等. 共情: 遗传-环境-内分泌-大脑机制. 科学通报, 2017, 62: 3729–3742]
27 Wang Y, Li W, Li C, et al. New insights into therapeutic strategies, drugs, and targets for advancing cancer therapy-related cardiovascular toxicity (in Chinese). Chin Sci Bull, 2025, 70: 991–993 [朱靖玮, 周雨青. 亲社会行为的内群体偏向: 社会动机的视角. 科学通报, 2025, 70: 991–1004]
28 Boyd R, Richerson P J. Culture and the evolution of human cooperation. Phil Trans R Soc B, 2009, 364: 3281–3288
29 Tomasello M. Why We Cooperate. Cambridge: MIT Press, 2009
30 Trivers R L. The evolution of reciprocal altruism. Q Rev Biol, 1971, 46: 35–57
31 Rand D G, Peysakhovich A, Kraft-Todd G T, et al. Social heuristics shape intuitive cooperation. Nat Commun, 2014, 5: 3677
32 Cialdini R B, Goldstein N J. Social influence: compliance and conformity. Annu Rev Psychol, 2004, 55: 591–621
33 House B R, Silk J B, Henrich J, et al. Ontogeny of prosocial behavior across diverse societies. Proc Natl Acad Sci USA, 2013, 110: 14586–14591
34 Nowak M A. Five rules for the evolution of cooperation. Science, 2006, 314: 1560–1563
35 Barclay P. Competitive helping increases with the size of biological markets and invades defection. J Theor Biol, 2011, 281: 47–55
36 Hruschka D J, Henrich J. Friendship, cliquishness, and the emergence of cooperation. J Theor Biol, 2006, 239: 1–15
37 Schino G, Aureli F. The relative roles of kinship and reciprocity in explaining primate altruism. Ecol Lett, 2010, 13: 45–50
38 Henrich J, Ensminger J, McElreath R, et al. Markets, religion, community size, and the evolution of fairness and punishment. Science, 2010, 327: 1480–1484
39 Dafoe A, Bachrach Y, Hadfield G, et al. Cooperative AI: machines must learn to find common ground. Nature, 2021, 593: 33–36
40 Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518: 529–533
41 Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. arXiv, 2017: 1707.06347
42 Sutton R S, Barto A G. Reinforcement Learning: An Introduction. 2nd ed. Cambridge: The MIT Press, 2018
43 de Curtò J, de Zarzà I. LLM-driven social influence for cooperative behavior in multi-agent systems. IEEE Access, 2025, 13: 44330–44342
44 Leibo J Z, Zambaldi V, Tacchetti A. Multi-agent reinforcement learning in sequential social dilemmas. In: Proc 16th Int Conf Auton Agents Multiagent Syst, 2017. 464–473
45 Lowe R, Wu Y, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments. In: Proc 31st Int Conf Neural Inf Process Syst, 2017. 6382–6393
46 Argyle L P, Busby E C, Fulda N, et al. Out of one, many: using language models to simulate human samples. Polit Anal, 2023, 31: 337–351
47 Park S A, Sestito M, Boorman E D, et al. Neural computations underlying strategic social decision-making in groups. Nat Commun, 2019, 10: 5287
48 Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017. 6000−6010
49 Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv, 2022: 2201.11903
50 Bertrand Q, Duque J, Calvano E, et al. Q-learners can provably collude in the iterated prisoner's dilemma. arXiv, 2023: 2312.08484
51 Lerer A, Peysakhovich A. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. arXiv, 2017: 1707.01068
52 Henrich J, Muthukrishna M. The origins and psychology of human cooperation. Annu Rev Psychol, 2021, 72: 207–240
53 Fehr E, Camerer C F. Social neuroeconomics: the neural circuitry of social preferences. Trends Cogn Sci, 2007, 11: 419–427
54 Gu R, He Y, Cui F. Reconsidering the relationship between empathy and prosocial behavior (in Chinese). Chin Sci Bull, 2025, 70: 982–990 [古若雷, 何越, 崔芳. 共情与亲社会行为的关系再思考. 科学通报, 2025, 70: 982–990]
55 Sampaio W M. The uniqueness of human cooperation. Nat Rev Psychol, 2024, 3: 70

56 Andreoni J, Bernheim B D. Social image and the 50–50 norm: a theoretical and experimental analysis of audience effects. Econometrica, 2009, 77: 1607–1636

57 Forsythe R, Horowitz J L, Savin N E, et al. Fairness in simple bargaining experiments. Games Economic Behav, 1994, 6: 347–369

58 Mengel F. Risk and temptation: a meta-study on prisoner's dilemma games. Economic J, 2018, 128: 3182–3209

59 De Cremer D, Tyler T R, Ouden N. Managing cooperation via procedural fairness: the mediating influence of self-other merging. J Economic Psychol, 2005, 26: 393–406

60 Tabibnia G, Lieberman M D. Fairness and cooperation are rewarding. Ann New York Acad Sci, 2007, 1118: 90–101

61 Zou X, Li D, Turel O, et al. Neural mechanisms of cooperation and fairness in iterative prisoner's dilemma. Behav Brain Res, 2025, 476: 115272

62 Declerck C, Boone C. Individual differences in prosocial decision making: social values as a compass. In: Declerck C, Boone C, eds. Neuroeconomics of Prosocial Behavior. Amsterdam: Elsevier, 2015. 111–145

63 Wichardt P C. Identity and why we cooperate with those we do. Behav Exp Econ, 2005, 29: 127–139

64 Mathew S, Boyd R. Punishment sustains large-scale cooperation in prestate warfare. Proc Natl Acad Sci USA, 2011, 108: 11375–11380

65 Herrmann B, Thöni C, Gächter S. Antisocial punishment across societies. Science, 2008, 319: 1362–1367

66 Baker B, Kanitscheider I, Markov T, et al. Emergent tool use from multi-agent autocurricula. 2019, arXiv: 1909.07528

67 Barton S L, Waytowich N R, Zaroukian E, et al. Measuring collaborative emergent behavior in multi-agent reinforcement learning. arXiv, 2018: 1807.08663

68 Brookins P, DeBacker J. Playing games with GPT: what can we learn about a large language model from canonical strategic games? Econ Bull, 2024, 44: 25–37

69 Engel C. Dictator games: a meta study. Exp econ, 2011, 14: 583–610

70 Jin S, Spadaro G, Balliet D. Institutions and cooperation: a meta-analysis of structural features in social dilemmas. J Pers Soc Psychol, 2025, 129: 286–312

71 Lorè N, Heydari B. Strategic behavior of large language models and the role of game structure versus contextual framing. Sci Rep, 2024, 14: 18490

72 Horton J J. Large language models as simulated economic agents: what can we learn from homo silicus? arXiv, 2023: 2301.07543

73 Phelps S, Russell Y I. The machine psychology of cooperation: can GPT models operationalise prompts for altruism, cooperation, competitiveness, and selfishness in economic games? 2023, arXiv: 2305.07970

74 Dal Bó P, Fréchette G R. Strategy choice in the infinitely repeated prisoner's dilemma. Am Econ Rev, 2019, 109: 3929–3952

75 Guo F. GPT in game theory experiments. 2023, arXiv: 2305.05516

76 Hosseini H, Khanna S. Distributive fairness in large language models: evaluating alignment with human values. 2025, arXiv : 2502.00313

77 Hackel L M, Doll B B, Amodio D M. Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. Nat Neurosci, 2015, 18: 1233–1235

78 Hackel L M, Zaki J. Propagation of economic inequality through reciprocity and reputation. Psychol Sci, 2018, 29: 604–613

79 Lockwood P L, Apps M A J, Valton V, et al. Neurocomputational mechanisms of prosocial learning and links to empathy. Proc Natl Acad Sci USA, 2016, 113: 9763–9768

80 Wittmann M K, Kolling N, Faber N S, et al. Self-other mergence in the frontal cortex during cooperation and competition. Neuron, 2016, 91: 482–493

81 Zhu L, Mathewson K E, Hsu M. Dissociable neural representations of reinforcement and belief prediction errors underlie strategic learning. Proc Natl Acad Sci USA, 2012, 109: 1419–1424

82 Li S, Chen X, Zhai Y, et al. The computational and neural substrates underlying social learning. Adv Psychol Sci, 2021, 29: 677–696

83 Rescorla R, Wagner A. A theory of pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black A H, Prokasy W F, eds. Classical Conditioning II: Current Research and Theory. New York: Appleton-Century-Crofts, 1972

84 Franklin N T, Frank M J. A cholinergic feedback circuit to regulate striatal population uncertainty and optimize reinforcement learning. eLife, 2015, 4: e12029

85 Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529: 484–489

86 Hadfield-Menell D, Dragan A, Abbeel P, et al. Cooperative inverse reinforcement learning. In: Lee D, Sugiyama M, Luxburg U, et al. eds. Advances in Neural Information Processing Systems. Vol. 29. London: Curran Associates, Inc., 2016

87 Li Y, Wei J, Liu Y, et al. Deep learning for micro-expression recognition: a survey. 2021, arXiv: 2107.02823

88 Wang Y, Li Y, Liang P P, et al. Cross-attention is not enough: incongruity-aware dynamic hierarchical fusion for multimodal affect recognition. 2023, arXiv: 2305.13583

89 Dell'acqua F, Kogut B, Perkowski P. Super Mario meets AI: experimental effects of automation and skills on team performance and coordination. Rev Econ Stat, 2025, 107: 951–966

90 Traeger M L, Strohkorb Sebo S, Jung M, et al. Vulnerable robots positively shape human conversational dynamics in a human–robot team. Proc

Natl Acad Sci USA, 2020, 117: 6370–6375
91 De Freitas J, Agarwal S, Schmitt B, et al. Psychological factors underlying attitudes toward AI tools. Nat Hum Behav, 2023, 7: 1845–1854
92 Bender EM, Gebru T, McMillan-Major A, et al. On the dangers of stochastic parrots: can language models be too big? FAccT′21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021. 610–623
93 Lazar S, Nelson A. AI safety on whose terms? Science, 2023, 381: 138
94 Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 2019, 366: 447–453
95 Schelble B, Flathmann C, Canonico L B, et al. Understanding human-AI cooperation through game theory and reinforcement learning models. In: Proceedings of the 54th Hawaii International Conference on System Sciences, 2021. 3–10

Summary for “人类与人工智能在社会互动中合作行为的异同与挑战”

Human–AI social cooperation: convergences, divergences, and challenges

Suchen Yao[1,2], Ji Shan[1,2], Li Hu[1,2] & Xuejing Lu[1,2*]

[1] *State Key Laboratory of Cognitive Science and Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China*
[2] *Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049, China*
* Corresponding author, E-mail: luxj@psych.ac.cn

Cooperation, a cornerstone of social interaction, not only shapes individuals’ behavioral decisions but also profoundly influences group cohesion and social evolution. From both evolutionary and cultural perspectives, humans demonstrate an “ultra-social” nature, characterized by consistent cooperation behaviors such as food sharing, caregiving, and mutual support observed across a wide range of ecological and societal contexts. From small foraging societies to large nations, individuals consistently demonstrate a strong willingness to cooperate with others. This cooperative capacity starts showing as early as middle childhood, highlighting deep evolutionary roots and critical adaptive advantages.

The rapid development of artificial intelligence (AI), particularly large language models (LLMs), has introduced a new generation of AI systems with remarkable capabilities in language comprehension, social reasoning, and contextual judgment, enabling them to participate in conversations, infer intentions, and respond to nuanced social cues in “human-like” ways. As AI becomes increasingly integrated into diverse social contexts and takes on more and more complex, varied roles, human–machine interactions are undergoing a profound transformation. This shift challenges our traditional theoretical frameworks of social interaction and raises urgent questions about the underlying mechanisms of cooperation when one or more partners are artificial agents.

To address this trend, this review systematically examines classic cooperation research and identifies three core stages of the cooperative process: strategy formation, behavior execution, and outcome evaluation and learning. Each stage encapsulates distinct cognitive operations and interaction dynamics. This continuous framework highlights that humans and AI not only exhibit distinct cognitive characteristics and behavioral modalities at each stage, but also manifest structural distinctions in inter-stage transitions. Within this framework, we summarize major findings from human interpersonal cooperation and human–AI cooperation. In addition to this, we compare significant differences in decision-making, cooperation patterns, and feedback adaptation between these two types of interaction. For example, although LLMs increasingly exhibit “human-like” cooperation tendencies, sometimes even outperforming human participants, these behaviors are driven by fundamentally different underlying processes. Moreover, when engaging with AI partners, humans often adjust their expectations differently than with human partners. This can lead to changes in trust formation, a slower development of shared routines, and an increased tendency to interpret AI behaviors through anthropomorphic lenses. Such patterns indicate cognitive and social adaptations unique to human-machine settings.

Based on these comparisons, we propose several strategies for optimizing AI systems in cooperative contexts, including enhancing the interpretability of AI decisions, designing more flexible real-time feedback mechanisms that respond to users, and ensuring transparent boundaries of agency to support smoother role negotiation between human and artificial agents. These insights enrich the current discourse on cooperation among both biological and artificial agents, providing practical guidance for advancing human-AI interaction models. By focusing on these areas, this review aims to deepen the understanding of cooperation between biological and artificial agents, and in doing so, it offers concrete suggestions for improving upcoming human-AI interaction models. It also underlines the crucial role of cognitive science and psychology in developing ethical, transparent, and socially responsible forms of cooperation in an increasingly blended human-AI society.

cooperation, social interaction, artificial intelligence (AI), human–machine cooperation, interpersonal cooperation

doi: 10.1360/CSB-2025-0662