人类亲社会性的策略塑造：一种潜在状态 POMDP 框架

Zahra Zahedi[1], Xinyue Hu[2], Shashank Mehrotra[1],Mark Steyvers[2],和 Kumar Akash[1]

arXiv:2603.02379v1 [cs.HC] 2 Mar 2026

摘要——**我们提出一种决策理论框架，使机器人能够在重复交互中策略性地塑造所推断的人类亲社会状态。将人类亲社会性建模为随时间演化的潜在状态，机器人学习通过自身动作（包括助人行为与信号传递）来推断并影响该状态。我们将其形式化为观测有限的潜在状态部分可观测马尔可夫决策过程，并使用期望最大化学习转移与观测动态。所得基于信念的政策在任务与社会目标之间取得平衡，选择能够最大化长期合作结果的动作。我们利用用户研究数据评估该模型，并表明学习到的策略在团队表现和提升所观测到的人类合作行为两方面均优于基线策略。**

索引术语——**认知建模、人类建模与仿真、人机协作**

一、引言

PROSOCIAL 行为是指自愿帮助他人的行为，在社会群体和合作系统的运作中发挥着核心作用。在日常互动中，人们常会遇到他人面临困境、苦苦挣扎或没有帮助便无法继续前进的情形。在此类情况下选择提供帮助，尤其是需要付出个人成本时，不仅体现了当下的任务目标，也反映出更深层的社会倾向。从教育到应急响应等各个领域，协作的形成与持续往往取决于人们何时、为何以及如何选择提供帮助。

亲社会性并非一种单一且始终如一的特质。行为科学区分出几种形式：主动型亲社会性是指在未受请求时主动提供帮助；反应型亲社会性是指在收到明确提示或请求后提供帮助；利他型亲社会性是指即使付出巨大成本或几乎不可能获得互惠，仍然提供帮助 [1]。其他形式还包括策略型亲社会性（通过提供帮助来促使他人日后提供帮助） [2]、受规范驱动的行为（通过提供帮助来遵循预期的社会角色）[3],以及共情关怀（基于共同情感而提供帮助）[4]。这些倾向因人、情境和时间而异，并且往往受到交互历史和对他人意图的感知影响。

稿件收到日期：2025年8月12日；修订日期：2025年11月16日；录用日期：2026年2月5日。本文经副编辑评估并参考审稿人意见后，由编辑 Angelika Peer 推荐发表。[1]Z. Zahedi、S. Mehrotra、K. Akash 任职于 Honda Research Institute USA, Inc. {zahra_zahedi, shashank_mehrotra, kakash}@honda-ri.com； [2]X. Hu、M. Steyvers 任职于加利福尼亚大学欧文分校认知科学系。 {xhu26, mark.steyvers}@uci.edu 通讯作者：Kumar Akash（kakash@honda-ri.com）。数字对象唯一标识符（DOI）：10.1109/LRA.2026.3668141。

在临时组队和松散协作的情境中，智能体追求各自的个体目标，但有时也可能彼此依赖；在这种情况下，亲社会行为尤为重要。由于智能体并未被明确要求相互帮助，合作必须自然而然地形成。智能体是否选择帮助另一个智能体，不仅取决于任务结构，也取决于其内部的亲社会状态，以及它如何解读他人过去的行为。这在人机交互中尤为重要，因为长期成功既取决于人类是否愿意合作，也取决于机器人能否通过自身动作影响这种意愿。

我们提出一种决策理论框架，使机器人能够随着时间推移策略性地塑造人类亲社会性。我们用 *亲社会塑造* 来指代这样一种行为：选择不仅影响即时结果，还能塑造机器人对人类潜在亲社会状态的信念。机器人与人类进行多轮互动，在每一轮中，双方都可能需要帮助。机器人不仅要决定是推进自己的任务，还是帮助对方，还要决定是否采取注意力引导行为（例如发出信号或突出显示），让自身需求更加显眼。这些选择会影响互动的未来走向，不仅通过即时结果发挥作用，还会影响对人类亲社会状态的推断；该状态与人类帮助他人的意愿有关，其中包括帮助机器人和其他人类。

我们将其形式化为一种潜在状态部分可观测马尔可夫决策过程（ls-POMDP）[1]。人类的亲社会状态无法直接观测，并会随着时间推移而变化，且受到机器人动作的影响。机器人接收部分且依赖于动作的反馈，例如观察人类在需要时是否提供帮助，并使用期望最大化（EM）方法，从记录的互动中学习潜在的转移和观测动态。由此得到的模型支持信念空间规划，兼顾任务进展和亲社会塑造这两个目标。

我们使用一项用户研究中收集的数据来评估这一方法。该研究在在线游戏环境中对反复进行的人机交互进行了考察。任务模拟了临时形成的相互依赖关系，在这种关系中，任一智能体都可能需要帮助才能继续进行。我们利用这些数据训练潜在状态模型，求解部分可观测马尔可夫决策过程以计算政策，然后在第二项涉及人类受试者的评估中，将所得行为与基线策略进行比较验证。我们的分析重点关注不同的机器人行为（例如助人行为与信号传递）如何影响人类对机器人的合作，并影响长期结果。

[1]虽然所有部分可观测马尔可夫决策过程根据定义都涉及潜在（未观测）状态，但我们使用“潜在状态部分可观测马尔可夫决策过程”这一术语，是为了强调此处所建模的人类亲社会状态作为一种不可观测成分，并非仅仅是环境性的，而是人类的一种结构化内在属性

本文已发表于IEEE机器人与自动化快报。此为作者版本。©2026 IEEE。保留所有权利，包括文本与数据挖掘以及人工智能及类似技术训练的权利。允许个人使用，但再版/再分发须获得IEEE许可。更多信息请参见https://www.ieee.org/publications/rights/index.html。

本研究的贡献如下：

- 一种用于在重复的临时人机交互中建模和塑造人类亲社会性的潜在状态部分可观测马尔可夫决策过程框架。该方法利用基于EM的学习，从部分可观测的动作-观测序列中估计亲社会状态动态，并通过信念空间规划，在互动约束下随时间影响人类的潜在行为。
- 利用两项用户研究的数据对模型和政策进行实证评估：一项用于学习潜在状态动态，另一项用于验证面向人类受试者的政策。结果表明，与基线策略相比，学习得到的政策能更有效地促进长期奖励和合作。

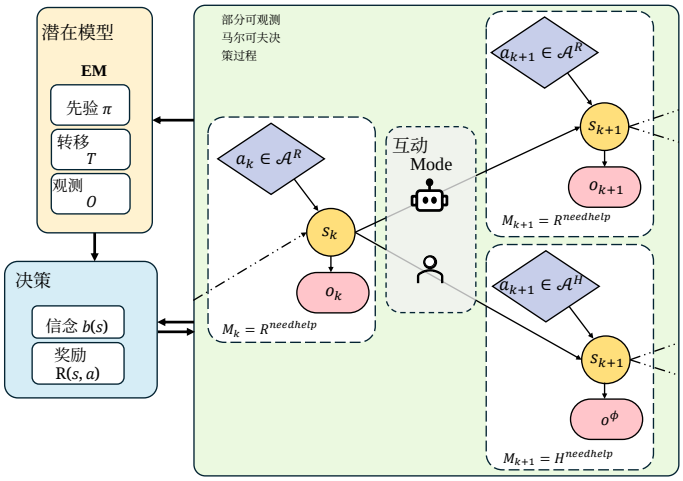


图 1：潜在状态部分可观测马尔可夫决策过程模型框架：机器人通过期望最大化学习亲社会状态动态，并利用信念空间决策来规划动作，以便在重复交互中策略性地塑造推断出的人类亲社会性。

II. 相关工作

亲社会行为包括合作、利他主义和互惠。社会互动的一个关键组成部分是互惠，即以善意回报善意 [5]。尽管人们与机器人合作的程度通常低于与人类合作的程度，但他们对两者表现出相似的互惠行为。这表明，互惠是一种影响互动的稳定特质 [5], [6]。近期模型将这些亲社会倾向视为潜在特质。例如， [7] 对潜在助人倾向进行建模，以改进对人类助人行为的预测。此类研究支持使用潜在状态模型来刻画人机交互中的亲社会性个体差异。

部分可观测马尔可夫决策过程常用于对信任、目标和偏好等潜在人类状态的不确定性进行建模。早期的人机交互研究通过人工方式编码人类行为，而近期研究则越来越多地从交互数据中学习这些潜在特质 [8]–[15]。例如， [11]提出了信任部分可观测马尔可夫决策过程，以动态推断人类信任并相应调整机器人行为； [16] 则对隐藏的人类子目标进行建模，以促进有效的双向沟通。同样，Ramachandran等人（2019年）采用辅助辅导部分可观测马尔可夫决策过程，根据推断出的参与状态个性化辅导 [14]。这些研究共同表明，基于信念的部分可观测马尔可夫决策过程框架能够在不确定情况下实现稳健决策。

当机器人通过审慎选择互动方式来策略性地影响人类行为时，这些建模方法的应用范围会进一步拓展 [17]–[19]。贝叶斯自适应模型和
动态贝叶斯网络能够有效降低对人类状态的不确定性，使机器人能够快速推断并增强合作与亲社会反应，尤其是在临时协作场景中 [9], [20], [21]。互惠会促使人类对机器人做出亲社会反应；值得注意的是，人类往往会回报非人类智能体所做的帮助行为，尤其是在这些行为得到明确传达时 [22], [23]。 [24] 强调，沟通策略能够显著改善合作动态。此外，机器人可以通过调整帮助程度或采用礼貌策略，有意影响人类信任与合作 [18], [25]。基于这些见解，我们提出一种潜在状态部分可观测马尔可夫决策过程方法，不仅能够推断
人类亲社会性，还能通过自适应决策策略性地塑造这种亲社会性，尤其适用于重复的临时人机交互。

III. 模型

A. 问题设定

我们考虑一个涉及两个交互智能体的序贯决策问题：机器人 R 和人类 H。每个智能体都独立执行个体任务。有时，由于环境约束（例如被阻挡或困住），某个智能体无法独自继续任务，因此可能需要另一个智能体的帮助。由于提供帮助的智能体必须暂时中断自己的任务，因此会承担成本。这就形成了即时个体目标与长期合作收益之间的权衡。

形式上，我们定义一个由 K 交互事件构成的离散序列。在每个交互事件 k中，一个智能体（R 或 H）可能需要另一个智能体的帮助，从而确定该事件的交互模式为 $M_k \in \{H^{\text{needhelp}}, R^{\text{needhelp}}\}$。机器人可以根据需要帮助的是哪个智能体而采取不同动作：(1) 如果人类需要帮助（$M_k = H^{\text{needhelp}}$），机器人可以选择提供帮助，即协助人类；也可以选择不提供帮助，继续执行自己的任务。(2) 如果机器人需要帮助（$M_k = R^{\text{needhelp}}$），机器人可以选择发出信号（例如突出显示并表明自己需要帮助），也可以选择不发出信号并保持沉默。虽然该框架可以扩展到更一般的情形和动作空间，但本文专注于对具体且基础的行为与动作进行建模。

因此，在每个交互事件 k 中，我们定义：

状态： 潜在亲社会状态 $s_k \in \mathcal{S}$，其中 $\mathcal{S} = \{s^0, s^1, \ldots, s^n\}$ 表示人类表现亲社会行为的倾向。请注意，我们的潜在状态代表人类一种结构化的内隐的内在亲社会倾向（而不仅仅是未被观察到的环境），即使人类自身也无法直接知晓。状态按从最低

本文已发表于 IEEE机器人与自动化快报。此版本为作者版本。©2026 IEEE。保留所有权利，包括文本与数据挖掘以及人工智能和类似技术训练方面的权利。允许个人使用，但重新发表/再分发须获得 IEEE 许可。更多信息请参见 https://www.ieee.org/publications/rights/index.html。

亲社会状态 s^0（例如，非亲社会）到最高亲社会状态 s^n 的顺序排列，亲社会倾向依次增强。

动作： 机器人动作 a_k 取决于当前互动M_k:

$$a_k \in \begin{cases} \mathcal{A}^H = \{a^{\text{help}}, a^{\text{no-help}}\}, & \text{if } \ M_k = H^{\text{needhelp}} \\ \mathcal{A}^R = \{a^{\text{signal}}, a^{\text{no-signal}}\}, & \text{if } \ M_k = R^{\text{needhelp}} \end{cases} \tag{1}$$

每种动作给机器人带来的成本各不相同： a^{help} 会产生帮助成本 C^{help}， a^{signal} 会产生信号成本 C^{signal}，而$a^{\text{no-help}}$ 和 $a^{\text{no-signal}}$ 不会产生即时动作成本。

观测： 只有当机器人明确请求帮助（M_k= R^{needhelp}）时，才能观测到人类响应（帮助 o^{help}/不提供帮助 $o^{\text{no-help}}$）；否则无法观测到该响应 $o^{\emptyset}$。事件 k的观测空间定义为 $o_k \in \mathcal{O} = \{o^{\text{help}}, o^{\text{no-help}}, o^{\emptyset}\}$。

由于观测会根据互动模式有条件地出现，我们将结构化的、依赖于模式的观测定义如下:

$$o_k \in \begin{cases} \{o^{\text{help}}, o^{\text{no-help}}\}, & \text{if } \ M_k = R^{\text{needhelp}} \\ \{o^{\emptyset}\}, & \text{if } \ M_k = H^{\text{needhelp}} \end{cases} \tag{2}$$

B. 模型表述

我们将该互动形式化为一个部分可观测马尔可夫决策过程（部分可观测马尔可夫决策过程），其结构围绕一系列交互事件 $k = \{1, \ldots, K\}$ 展开。我们将该部分可观测马尔可夫决策过程定义为一个元组 $\mathcal{M}$ =($\mathcal{S}$, $\mathcal{A}$, $\mathcal{O}$, T, O, R, γ)，其中每个组成部分的正式定义如下:

转移模型： 交互事件之间的潜在状态转移明确取决于当前状态和机器人选择的动作:

$$T(s_{k+1}|s_k, a_k) = P(s_{k+1}|s_k, a_k), \quad s_k, s_{k+1} \in \mathcal{S}. \tag{3}$$

观测模型： 条件观测概率明确取决于交互模式和潜在亲社会状态:

$$O(o_k|s_k, a_k, M_k) = \begin{cases} P(o_k|s_k, a_k), & \text{if } \ M_k = R^{\text{needhelp}} \\ \delta_{o_k = o^{\emptyset}}, & \text{if } \ M_k = H^{\text{needhelp}} \end{cases} \tag{4}$$

这里， δ 是克罗内克δ函数，因为 $M_k = H^{\text{needhelp}}$始终不会产生观测 $o^{\emptyset}$。

奖励模型： 每个交互事件 k 的奖励旨在反映助人行为和信号传递的即时成本，以及受潜在亲社会状态影响的长期合作收益:

$$\begin{aligned} R(s_k, a_k) = &- C^{\text{help}}\mathbb{I}(a_k = a^{\text{help}}) - C^{\text{signal}}\mathbb{I}(a_k = a^{\text{signal}}) \\ &+ R^{\text{prosocial}}(s_k), \end{aligned} \tag{5}$$

其中， $R^{\text{prosocial}}(s_k)$表示与人类更高亲社会状态相关的间接合作与社会效益。模型示意图见图1。

C. 模型学习

与标准的部分可观测马尔可夫决策过程（POMDP）表述不同，标准表述中的转移模型和观测模型通常根据完全观测到的状态-动作序列指定或学习，而我们的潜在状态部分可观测马尔可夫决策过程模型涉及一种无法直接观测的潜在亲社会状态。因此，必须从动作-观测轨迹中推断状态。这并不会改变马尔可夫假设；转移在潜在状态和动作上仍是一阶的。因此，我们使用期望最大化（EM）算法，具体来说是适用于部分可观测马尔可夫决策过程的 Baum-Welch变体 [26], [27],，根据部分可观测交互数据估计转移参数和观测参数。我们使用期望最大化算法根据记录的交互序列估计模型参数:

$$\mathcal{D} = \{(M_k, a_k, o_k)\}_{k=1}^{K} \tag{6}$$

在*E*步中，我们根据潜在亲社会状态计算前向-后向概率，以估计预期状态占用率和状态转移，并纳入动作条件观测。此步骤遵循标准的 Baum-Welch 前向-后向消息计算方法。

在*M*步中，我们根据 E 步得到的潜在状态分布估计值，通过最大化期望似然来更新转移参数和观测参数。标准的期望最大化公式假定每一步都有观测，而我们的模型则明确处理了结构化、依赖模式的可观测性:

转移更新（处理无观测情形）： 对于没有可观测人类响应的交互事件（即$M_{k+1} = H^{\text{needhelp}}$），计算预期状态转移时不纳入观测似然。具体而言，此类事件的前向-后向更新仅涉及转移模型和后向消息，不包含观测似然的贡献。给定动作 a_k时，从状态 s_k 转移到 s_{k+1}的预期转移概率计算如下:

$$\xi_k(s_k, s_{k+1}, a_k) = b_k(s_k)\, T(s_{k+1} \mid s_k, a_k)\, \frac{\beta_{k+1}(s_{k+1})}{\beta_k(s_k)} \tag{7}$$

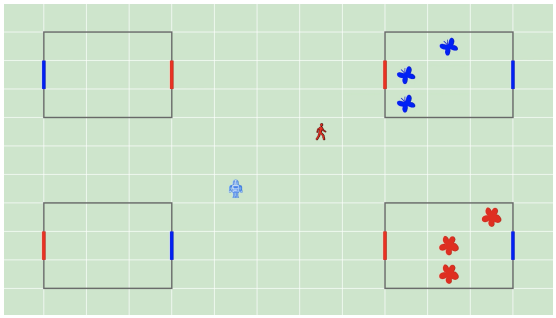
观测更新（仅适用于可观测情形）： 观测参数仅根据人类响应明确可观测的交互事件（即$M_k = R^{\text{needhelp}}$）进行更新。对于没有可观测响应的交互事件，不更新观测模型。

通过明确处理这些依赖模式的情形，我们的基于EM的学习方法能够从部分可观测的交互序列中估计潜在状态转移和观测参数。

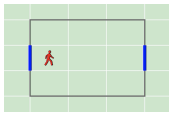
D. 规划与决策

在规划阶段，机器人会显式维护并推理信念状态 $b_k(s_k)$，以表示其对交互事件k中人类潜在亲社会状态的不确定性。利用学习到的模型参数，我们在信念空间中进行规划，以求解由此产生的部分可观测马尔可夫决策过程决策问题。在每个交互事件中，机器人会策略性地选择动作 a_k ，以最大化预期累积

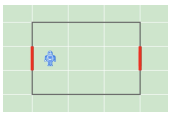
本文已发表于《IEEE机器人与自动化快报》。此版本为作者版本。©2026 IEEE。保留所有权利，包括文本与数据挖掘以及人工智能和类似技术训练方面的权利。允许个人使用，但再版或再分发须获得 IEEE 许可。更多信息请参阅 https://www.ieee.org/publications/rights/index.html。



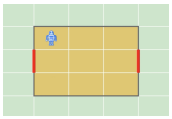
(a) 游戏环境概览



(b) $H^{needhelp}$



(c) $R^{needhelp}$



(d) a^{signal}

图2：游戏设置概览、不同模式 M_k以及作为机器人动作的信号传递。

折扣奖励，同时显式考虑即时结果以及其动作对人类潜在亲社会倾向的长期影响。形式化地说，我们在信念空间中求解以下贝尔曼方程：

$$V^*(b_k) = \max_{a_k \in \mathcal{A}_{M_k}} \Big[\sum_{s_k \in \mathcal{S}} b_k(s_k) R(s_k, a_k) + \gamma \sum_{o_k \in \mathcal{O}} P(o_k | b_k, a_k) V^*(b_{k+1}) \Big] \quad (8)$$

这里， $R(s_k, a_k)$表示即时成本（例如助人行为、信号传递）以及与影响人类亲社会行为相关的间接收益。折扣因子 $\gamma \in [0, 1)$ 用于权衡即时奖励与未来奖励。

每次交互事件发生后，信念状态 $b_k(s_k)$ 都会根据标准的 POMDP信念更新规则递归更新，并纳入观测到的人类响应 o_k、所选动作 a_k 和交互模式 M_k。这确保机器人能够系统地纳入观测结果和潜在状态的不确定性，从而在重复交互中策略性地塑造人类亲社会性。

IV. 实验

A. 用户研究

为了评估我们的模型和规划框架，我们利用了一项先前发表的用户研究中的数据 [22]，并通过新增两种实验条件（不提供帮助/有信号和不提供帮助/无信号）对其进行了扩展。完整数据集包含 540 名人类参与者与机器人进行五轮代币收集游戏的互动。该研究采用重复交互结构，旨在模拟非对称临时相互依赖。研究的目标是在受控的合作设置下收集交互轨迹，以便进行模型学习和评估。该研究已获加利福尼亚大学欧文分校机构审查委员会（IRB）批准。

实验设计：参与者与机器人在二维网格世界环境中进行五轮代币收集游戏。研究明确告知人类玩家，他们将与机器人一起游戏。每位玩家（人类和机器人）收集出现在不同房间中的不同代币。玩家收集完自己的代币后，另一间房间中会随机出现一组三个新代币。游戏没有明确说明这种互动是协作还是竞争。每轮开始时，人类和机器人分别位于网格的对角位置，且无法收集对方的代币。机器人采用A^* 寻路算法实现，以每秒两个网格单元的恒定速度移动。

每个房间都有一扇红门和一扇蓝门，分别对应人类玩家和机器人玩家。进入房间后，房间门的颜色会互换。如果两扇门都变成另一位玩家的颜色，房间内的玩家就会被困住，无法离开（见图2b和图2c）。只有另一位玩家选择解救被困者，解锁房门，才能将其释放；解锁房门的方式是进入被困玩家所在的房间。每回合中，人类或机器人会被困住（M_k），需要帮助才能继续，而另一位玩家可以进入房间将其解救。为统一所有参与者的时间安排，每回合开始20秒后，玩家会在四个房间之一被困。图2展示了代币收集游戏的设置、两位玩家被困时的示例情形以及信号传递。帮助机会的时间安排和结构遵循六种固定的回合级条件，用以控制每位玩家何时被困。这六种条件旨在确保两位玩家都有多次机会展现亲社会行为。在此情境中，“H”表示人类有机会帮助机器人的回合，“R”表示机器人有机会帮助人类的回合。六种帮助机会序列为 RRHHH、HRRHH、HHRRH、RHRHH、HRHRH、RHHRH。每位参与者会被随机分配到其中一种序列。有关游戏设计和研究流程的更多信息，请参阅补充视频。数据集可通过此链接获取。

在被试间设计中，机器人行为还因四种实验条件而异：2×2

- 提供帮助（$a_{\text{帮助}}$）与不提供帮助（$a_{\text{不提供帮助}}$）：当人类被困时，机器人是否提供帮助；如果提供帮助，机器人会在被困事件发生5秒后开始朝他们移动。• 发出信号（$a_{\text{发出信号}}$）与不发出信号（$a_{\text{不发出信号}}$）：当人类被困时，机器人是否利用视觉显著性（例如让房间闪烁）来吸引人类的注意力（图2d）。

人类受试者：所有参与者均通过 Prolific 在线招募。所有参与者均已提供知情同意，且未接受任何关于帮助机器人的指示或激励，因此能够自然地表现出助人行为。在 540 名参与者中，308 人自我认同为男性，232 人自我认同为女性，年龄范围为 19 至 72 岁（M = 37，SD = 12）。无论表现如何，参与者均获得固定报酬 3.0 美元。他们均自我报告称，在实验进行时已年满 18 岁，且为居住在

本文已发表于《IEEE机器人与自动化快报》。这是作者版本。©2026 IEEE。保留所有权利，包括文本与数据挖掘以及训练人工智能和类似技术的权利。允许个人使用，但转载/再分发须经 IEEE 许可。更多信息请参阅 https://www.ieee.org/publications/rights/index.html。

美国的英语使用者。四种条件下的参与者人数分别为 提供帮助 + 无信号 ($n = 177$)、提供帮助 + 信号($n = 164$)、不提供帮助 + 信号 ($n = 104$)，以及 不提供帮助 + 无信号 ($n = 95$)。使用 G–Power [28]； $f = 0.25$； $\alpha = 0.05$； $1 - \beta = 0.95$) 进行了先验功效分析，结果表明$N \geq 200$，因此 $N = 540$ 对中等效应具有足够的效力。

统计分析：为测量人类助人行为的可能性（o^{help}），研究采用了使用 logit 链接函数的广义混合效应模型。信号（a^{signal}）和机器人帮助（a^{help}）的实验条件被视为固定预测变量。与零模型相比，机器人帮助显著改善了模型拟合度（$\Delta AIC = -40.2$），$\chi^2(1) = 43.87, p < 0.001$。模型方程如下：

$$\log\left[\Pr(o^{\text{help}} = 1)\right] = \beta_0 + \beta_1\, a^{\text{signal}} + \beta_2\, a^{\text{help}} + (1|Subject) \tag{9}$$

其中，β_i 为模型系数，(1|Subject)表示每位参与者的随机截距。研究结果表明，机器人助人行为（a^{help}）提高了人类助人行为的优势（o^{help}）（优势比 = 10.42，95% CI [5.09, 21.32]，$z = 6.44$，$p < 0.001$）。这意味着，当机器人此前提供过帮助时，参与者提供帮助的可能性是原来的 10.42 倍。信号传递 a^{signal} 也呈现上升趋势，但未达到统计显著性（优势比 = 1.74，95% CI [0.94, 3.22]，$z = 1.77$，$p < 0.077$）。这一对比表明，助人行为可能在不同亲社会状态下具有更广泛的强化作用，而信号传递的影响可能更容易受到具体潜在亲社会状态的影响；例如，它在某些状态下（如亲社会性较高时）可能有效，在其他状态下则效果较弱。固定效应模型无法捕捉这些模式，但我们的潜在状态框架可以对此加以解释：信号传递以依赖信念的方式影响亲社会状态之间的转移概率，并根据对人类亲社会状态的信念选择动作，因此仍被视为重要的动作变量。此外，实验条件之间未观察到显著的交互效应，表明条件分配并未使结果产生偏差。

B. 模型实现

对于每位参与者，我们构建了一条涵盖五轮交互轮次的交互轨迹。在每一回合，我们记录了交互模式 M_k （即哪一个智能体被困住）、机器人所选的动作 a_k，以及人类的观测-

如果机器人被困，则会作出可行的响应 o_k 。这些轨迹构成了一个数据集，用于学习潜在状态部分可观测马尔可夫决策过程的转移参数和观测参数，并评估策略在模拟条件和留出条件下的性能。

模型训练：我们使用记录的互动序列来训练转移模型 $T(s_{k+1} \mid s_k, a_k)$、观测模型 $O(o_k \mid s_k, a_k, M_k)$ 和初始潜在状态分布 $b_0(s)$。期望最大化算法采用转移模型和观测模型的对称狄利克雷先验 ($\alpha = 1$) 以及均匀初始信念 $b_0(s)$ 进行初始化。我们使用对数似然评估了包含 2 至 5 个隐状态的模型。每种配置均进行了 30 次随机重启训练。4 状态模型的对数似然最高

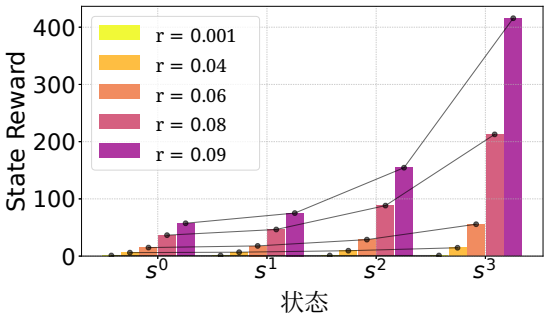


图 3：不同奖励梯度 r 下的状态奖励值，展示了在不同设定下，奖励如何随着亲社会性的提高而增加。

模型 (−560.23)。据此，我们选择了包含四种状态的离散潜在状态空间 ($|\mathcal{S}| = 4$)，该空间最能捕捉从参与者交互中推断出的亲社会行为模式差异。训练完成后，我们将学得的状态转移动态和潜在状态结构（例如，状态连通图和转移概率）可视化并加以分析，以便根据推断出的亲社会水平解读并标注各状态。基于转移结构和稳定性，我们识别出亲社会性较高的状态，以及反映较低或中等水平的状态。四种潜在状态分别标记为 s^0：非亲社会、s^1：低亲社会、s^2：中等亲社会，以及s^3：高亲社会。

奖励与成本设定：根据观测到的行为数据，我们明确设定了机器人动作成本和状态相关奖励。机器人动作成本依据交互研究中观察到的大致打断成本设定。具体而言，涉及直接协助(a_{help})的动作会产生与机器人帮助人类时估计的平均打断时间成正比的成本$C^{\text{help}}\mathbb{I}(a_k = a^{\text{help}}) = 15$。信号传递($a^{\text{signal}}$)的成本设定为与助人行为相同，以反映相近的打断程度；敏感性分析纳入了不同的变化，而不采取动作的选择($a_{\text{no-help}}$，$a_{\text{no-signal}}$)则成本为零。

状态相关奖励旨在体现亲社会性逐步提高的程度，其设计参考了不同交互条件下代币收集表现所呈现的趋势。我们将状态奖励参数化如下：

$$R^{prosocial}(s) = \exp(r \cdot x_s) - C, \tag{10}$$

其中，x_s 表示与潜在状态 s相关联的亲社会得分，r 控制奖励梯度的陡峭程度。常数 $C = \exp(r \cdot x_{s^0})$用于将最低亲社会状态的奖励归一化至接近零。 x_s的取值基于与每个推断潜在状态相关的代币总收集表现的实证观察，并设定为：$x_{s^0} = 45$、$x_{s^1} = 48$、$x_{s^2} = 56$和 $x_{s^3} = 67$。这一表述方式可根据设计者的目标和情境激励，灵活控制奖励结构。通过调整指数参数r，模型可以表示从各状态间奖励几乎均匀，到奖励急剧增加并更加重视较高亲社会性的广泛条件。图

本文已发表于 IEEE机器人与自动化快报。这是作者版本。©2026 IEEE。保留所有权利，包括文本与数据挖掘以及训练人工智能和类似技术的相关权利。允许个人使用，但转载/再分发须获得 IEEE 许可。更多信息请参见 https://www.ieee.org/publications/rights/index.html。

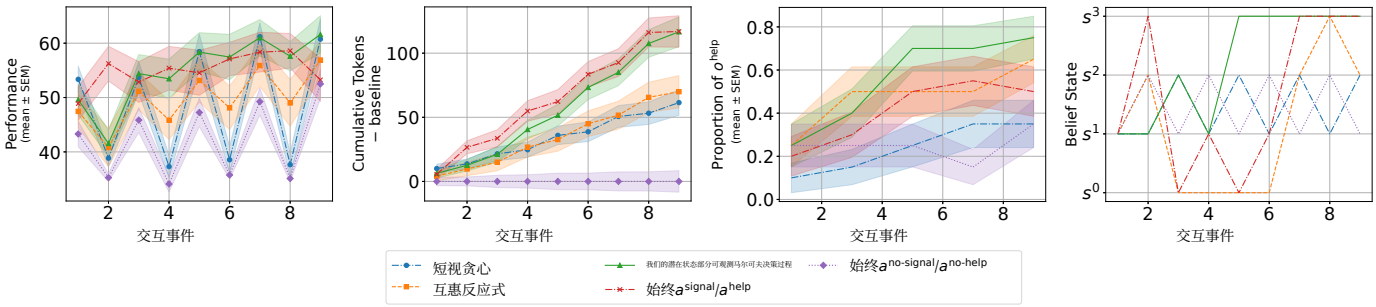


图4：9个交互步骤中的政策比较。在累计代币方面，将从不帮助／从不发信号视为基线因素。

3展示了改变奖励指数 r 如何影响各状态间的奖励分布。敏感性分析部分探讨了这些不同的指数取值如何影响学习到的策略。

决策与策略求解：利用学习到的模型参数，我们将机器人决策形式化为一个信念空间规划问题。在每个交互事件中，机器人都会维护信念状态 $b_k(s_k)$，表示对人类潜在亲社会状态的不确定性，并根据观测事件更新这一信念。为了选择能够在多次互动中策略性塑造亲社会行为的最优动作，我们使用 SARSOP 求解器 [29], 求解信念空间部分可观测马尔可夫决策过程；这是一种用于近似求解部分可观测马尔可夫决策过程的高效算法。计算得出的政策明确考虑了即时成本、依赖状态的亲社会奖励，以及机器人动作对人类行为的长期影响。

V. 评估

A. 策略评估与基线比较

为了评估我们学习得到的潜在状态部分可观测马尔可夫决策过程策略在行为场景中随时间的表现，我们采用第IV-A节所述的同一款代币收集游戏开展了一项新的人类受试者研究，并将游戏扩展为九个交替模式的交互轮次 M_k of $R^{needhelp}$ and $H^{needhelp}$ 条件（*HRHRHRHRH*）。这种设置确保了在整个互动过程中，两个智能体都有平等的机会去帮助对方或接受对方的帮助。

我们将自己的策略与四种基线策略进行了比较：1）**总是帮助／总是发信号：** 机器人始终选择动作（a^{help}, a^{signal}）。2）**从不帮助／从不发信号：** 机器人始终选择无动作选项（$a^{\text{no-help}}$, $a^{\text{no-signal}}$）。3）**短视贪心策略：** 一种单步前瞻策略，选择该动作。

$$a_k^* = \arg\max_{a_k \in \mathcal{A}_{M_k}} \underbrace{\sum_{s_k} b_k(s_k)\, R(s_k, a_k)}_{\text{immediate expected reward}} + \underbrace{\sum_{s_k, s_{k+1}} b_k(s_k)\, T(s_k, a_k, s_{k+1})\, R^{prosocial}(s_{k+1})}_{\text{next-state (lookahead) value}}, \tag{11}$$

最大化即时期望奖励，而不考虑未来的信念演化。

表 I：混合方差分析结果 st

	团队表现		观测		累计代币	
	F	p值	F	p值	F	p值
政策 (P) (4)	9.8314	0.0000*	2.8911	0.0263*	4.6720	0.0017*
时间 (T) (1)	50.5082	0.0000*	33.1379	0.0000*	1907.6132	0.0000*
P× T (4)	1.1264	0.35	1.5862	0.18	5.0390	0.0010*

表 II：采用 Tukey 法的两两比较 HSD.

组 – 潜在状态部分可观测马尔可夫决策过程	团队表现		观测	
	Δ	p值调整	Δ	p值调整
始终帮助/发信号	0.027	1.0	−0.08	0.29
从不帮助/发信号	−12.96	0.0*	−0.17	0.0006*
反应式	−5.19	0.01*	−0.044	0.84
短视	−6.13	0.001*	−0.18	0.0003*

4) 互惠反应式： 一种反应式策略：如果人类在上一次互动中提供了帮助，机器人就会提供帮助（a^{help}），否则不提供帮助（$a^{\text{no-help}}$）；在信号传递方面，如果人类在上一次互动中没有提供帮助，机器人就会发送信号（a^{signal}），否则不发送信号（$a^{\text{no-signal}}$）。

对于基于奖励的策略（潜在状态部分可观测马尔可夫决策过程 和短视贪心策略），奖励指数参数设为 $r = 0.06$，从而在亲社会状态上提供平衡且逐步递增的奖励。

在与第IV-A节研究相同的功效分析假设下，结合本研究的混合设计规范（策略 × 回合），分析得出的最低所需样本量为 $N \geq 90$。因此，我们从 Prolific 招募了 $N = 100$ 名参与者（在五种策略条件下保持均衡；年龄：$M = 41.5$, SD =12.08；女性占 53%，男性占 46%）。无论表现如何，参与者均获得 4 美元的固定报酬。

我们从以下方面评估了每种策略的表现：(a)团队表现：双方智能体每回合收集的代币总数，(b) 累计代币：累积至每一回合的代币总数，(c) 信念状态演变：机器人对人类亲社会状态所持信念的演变，以及(d)观察到的人类助人行为 o^{help} 比率：观察到帮助 o^{help} 的参与者在每一 R^{needhelp} 回合中所占的比例。

如图 4 所示，我们的政策在所有指标上均优于所有基线，包括团队表现、累计代币、观察到人类帮助的比例以及信念状态的变化。虽然“始终帮助／始终发出信号”政策的团队表现相近，但我们的潜在状态部分可观测马尔可夫决策过程政策所需的高成本行动更少，观察到的人类帮助比例也更高

本文已发表于IEEE机器人与自动化快报。此为作者版本。©2026 IEEE。保留所有权利，包括文本与数据挖掘以及人工智能及类似技术训练的权利。允许个人使用，但再版/再分发需获得IEEE许可。更多信息请参见https://www.ieee.org/publications/rights/index.html。

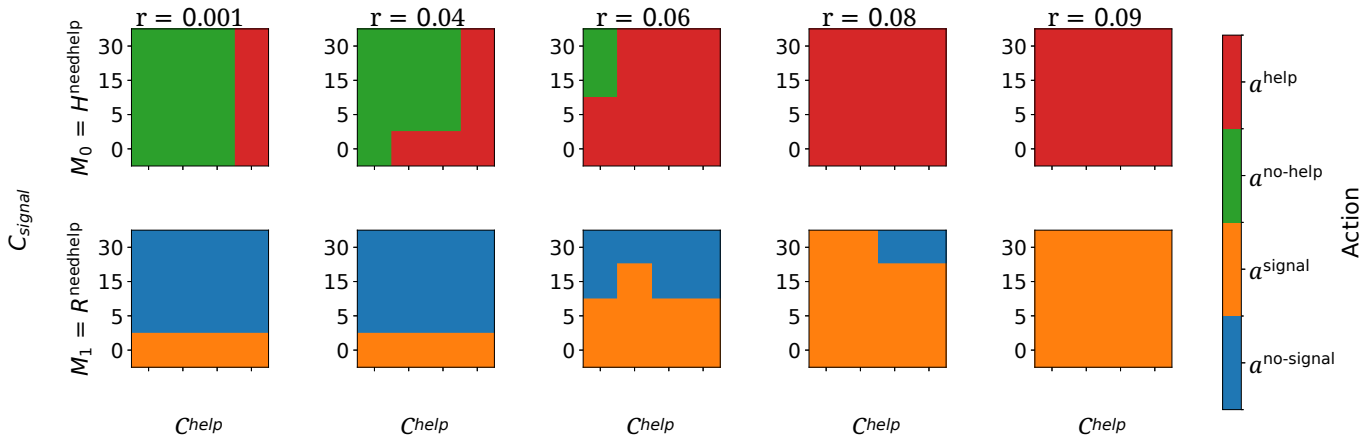


图 5：机器人策略对奖励和成本参数的敏感性。

。这种相近的表现符合预期，因为团队绩效指标基于代币，且不对机器人动作成本进行惩罚，因此在此任务中，助人行为可以使表现达到饱和。当较高亲社会状态带来的奖励超过行动成本时，潜在状态部分可观测马尔可夫决策过程的行为与始终帮助相似；在其他设置下，它会有选择地不采取高成本行动，这与固定基线不同。这一优势源自信念感知规划，使机器人能够采取成本效益高的行动，同时不损害其长期影响力。与“始终帮助／始终发出信号”条件相比，机器人对人类潜在亲社会状态的信念也更快地向较高亲社会水平移动，表明其推断出的亲社会塑造过程更快；这与观察到的人类帮助率较高（o^{help}）一致，并证实机器人成功促成了合作。

总的来说，这些结果表明，我们的政策不仅能够高效地实现更高的团队表现，还能比基线更有效、更快速地提升观测到的助人行为，在短时互动中同时实现更高的表现和更快的塑造。

我们采用混合设计方差分析，评估政策和时间（开始交互事件与最终交互事件）对团队表现、累计代币和人类帮助观测的影响。结果显示，政策和时间均具有显著主效应（均为 $p < 0.02$）。此外，累计代币在政策与时间之间表现出显著交互作用（$p < 0.001$），表明不同政策不仅在累计表现上存在差异，其行为随时间演变的方式也不同（见表 I）。

如表 II 所示，针对两两比较的 Tukey诚实显著差异检验事后检验进一步表明，与除总是帮助／总是发信号之外的所有政策相比，我们的潜在状态部分可观测马尔可夫决策过程政策在团队表现方面显著更优（$p < 0.01$）。就观察到的人类帮助而言（ o^{help}），我们的政策显著高于（$p < 0.001$）从不帮助／从不发信号和短视贪心策略。

B. 敏感性分析

我们进行了系统的敏感性分析，以评估奖励梯度和行动成本的变化如何影响机器人的决策策略。具体而言，我们考察了机器人选择的动作如何随以下参数取值范围的变化而改变：（1）**奖励缩放因子** $r \in \{0.001, 0.04, 0.06, 0.08, 0.09\}$，该参数控制指数奖励函数中的奖励随亲社会性增加而上升的陡峭程度。（2）**行动成本**C_{help}，$C_{\text{signal}} \in \{30, 15, 5, 0\}$，这些成本会惩罚助人行为和信号行为。

为评估所得政策，我们模拟了一个两步交互场景：$H^{\text{needhelp}} \rightarrow R^{\text{needhelp}}$，并交替改变被困的智能体。我们将从数据中学习到的先验信念用作初始信念。对于每种（r，C_{help}，C_{signal}）配置，我们求解部分可观测马尔可夫决策过程，并记录机器人在每个时间步的动作。图 5 所示的结果呈现出明显不同的行为模式：

- 在首次互动中，$k = 0$（$M_0 = H_{\text{需要帮助}}$），只有当 $r \geq 0.08$时，帮助才会成为主导选择，这反映出需要更强的激励才能证明付出代价的亲社会行为是合理的。当助人成本较高或奖励差异较小时，机器人会默认采取 $a_{\text{不提供帮助}}$。
- 在上一次互动中，$k = 1$ （$M_1 = R_{\text{需要帮助}}$），在先前的助人动作之后，机器人更有可能进行信号传递，尤其是当 $r \geq 0.06$ 且信号成本为中等或较低时。随着奖励梯度变陡，即使成本较高，机器人也愿意进行信号传递

这些发现证实，机器人的政策会适应不同的激励结构，包括奖励梯度与行动成本的差异。在 r较低的情形下，机器人采取保守策略，避免产生成本的动作。随着奖励梯度变陡，政策会转向更多促进亲社会行为的行动，尤其是在成本适中时。这凸显了该模型在适应不同规范性假设以及成本与长期奖励之间的权衡方面的灵活性。

六、讨论与结论

我们的评估表明，学习得到的潜在状态部分可观测马尔可夫决策过程政策，能够在重复交互中实现基于信念、具有策略适应性的机器人行为。机器人根据其对人类亲社会性的内部信念选择动作。

本文已发表于IEEE机器人与自动化快报。此为作者版本。©2026 IEEE。保留所有权利，包括文本与数据挖掘以及人工智能及类似技术训练的权利。允许个人使用，但再版/再分发需获得IEEE许可。更多信息请参见https://www.ieee.org/publications/rights/index.html。

结果证实，我们的政策在团队表现和人类合作行为方面均优于基线，成功引导互动朝着更高的观测助人行为水平和推断出的亲社会状态发展。信念轨迹表明，机器人不仅会对观测到的行为做出反应，还会主动影响未来的合作。需要指出的是，尽管我们将亲社会性建模为一种潜在内部状态，但在本研究中，人类的真实意图无法直接观测，也未通过实验加以控制。因此，我们的结论基于推断出的行为模式，而非经验证的内在动机。

我们的分析表明，与助人行为相比，信号传递对状态和成本更为敏感。当亲社会塑造的奖励很高且成本适中时，机器人会选择性地进行信号传递；而当亲社会状态受到重视且成本可控时，助人行为的使用则更为广泛。敏感性分析证实了这一区别，表明信号传递行为会随参数变化而急剧变化，而当亲社会状态受到高度重视时，助人行为则更稳定地受到偏好。

对人类亲社会性进行建模与塑造的能力，与那些合作必须自然涌现而非通过明确指令产生的领域直接相关。这包括灾害应对或太空任务中的临时团队协作、共享自主系统中的同伴支持、教育技术中的社会学习，以及服务机器人在公共环境中的自发合作。在这些情境中，通过重复交互塑造推断出的潜在亲社会状态，可能使机器人在没有预定义协调协议的情况下促进更可靠的长期合作。

未来工作应探索更一般的动作设定、个性化建模，以及基于长期行为数据的奖励塑造。此外，尽管本研究聚焦于受控任务中的短期互动，将其扩展到更复杂且开放的领域仍是一个有待解决的挑战。

最后，塑造人类行为会引发重要的伦理问题。部署保障措施包括：对适应过程保持透明、获得用户同意/提供退出选择、采取最低必要程度的影响、设定有界的行动成本，以及保留审计日志。未来的工作应考察透明度、同意以及用户对机器人发起的影响的看法，旨在制定符合伦理要求的亲社会塑造方法。

总而言之，我们的工作为人机交互中的社会适应性决策提供了一个基于原则的框架，使机器人能够在重复的临时互动中对人类亲社会性进行建模并施加影响。

参考文献

[1] D. Findley-Van Nostrand 和 T. Ojanen，《亲社会行为形式与青少年的社会目标及同伴地位存在差异化关联》，遗传心理学杂志，第 179 卷，第 6 期，第 329–342 页，2018年。[2] S. Grueneisen 和 F. Warneken，《亲社会行为的发展——从同情到策略》，心理学最新观点，第 43 卷，第 323–328 页，2022年。[3] B. R. House、J. Henrich、S. F. Brosnan 和 J. B. Silk，《人类亲社会性的发生：对 3 至 8 岁儿童开展的行为实验》，进化与人类行为，第 33 卷，第 4 期，第 291–308 页，2012年。[4] C. D. Batson，《人类利他主义》。牛津大学出版社，2011年。[5] E. B. Sandoval、J. Brandstetter、M. Obaid 和 C. Bartneck，《人机交互中的互惠：通过囚徒困境和最后通牒博弈开展的定量研究》，《国际社会机器人学杂志》，第 8 卷，第 303–317 页，2016年。

[6] A. Mahmoodi、B. Bahrami 和 C. Mehring，“社会影响的互惠性”，《自然·通讯》，第9卷第1期，第2474页，2018年。[7] A. Nanavati、C. I. Mavrogiannis、K. Weatherwax、L. Takayama、M. Cakmak 和 S. S. Srinivasa，“结合个体和情境因素对人类助人行为进行建模，以用于机器人规划”，载于《机器人：科学与系统》，2021年，第1–11页。[8] S. Nikolaidis、R. Ramakrishnan、K. Gu 和 J. Shah，“从联合行动示范中高效学习模型，用于人机协作任务”，载于第十届ACM/IEEE人机交互国际会议论文集，2015年，第189–196页。[9] S. Nikolaidis、S. Nath、A. D. Procaccia 和 S. Srinivasa，“人机协作中人类适应的博弈论建模”，载于2017年ACM/IEEE人机交互国际会议论文集，2017年，第323–331页。[10] M. Natarajan、C. Xue 和 M. Gombolay，“非最优性条件下的混合主动式人机团队协作”，载于2023年AAAI秋季研讨会系列，2023年。[11] M. Chen、S. Nikolaidis、H. Soh、D. Hsu 和 S. Srinivasa，“人机协作中的信任感知决策：模型学习与规划”，《ACM人机交互汇刊》（THRI），第9卷第2期，第1–23页，2020年。[12] Z. Zahedi、S. Sreedharan、M. Verma 和 S. Kambhampati，“对人类信任与监控之间相互作用的建模”，载于2022年第17届ACM/IEEE人机交互国际会议（HRI）。IEEE，2022年，第1119–1123页。[13] J. Lee、J. Fong、B. C. Kok 和 H. Soh，“彼此了解：为人机协作校准意图、能力和信任”，载于2020年IEEE/RSJ智能机器人与系统国际会议（IROS）。IEEE，2020年，第6296–6303页。[14] A. Ramachandran、S. S. Sebo 和 B. Scassellati，“使用辅助导师部分可观测马尔可夫决策过程（AT-POMDP）进行个性化机器人辅导”，载于AAAI人工智能会议论文集，第33卷第01期，2019年，第8050–8057页。[15] Z. Zahedi、S. Mehrotra、T. Misu 和 K. Akash，“迈向知情的自动驾驶汽车决策：出行中的福祉与信任计算模型”，IJCAI，2025年。[16] V. V. Unhelkar、S. Li 和 J. A. Shah，“序贯人机协作任务中双向沟通的决策”，载于2020年ACM/IEEE人机交互国际会议论文集，2020年，第329–341页。[17] S. Sheng、P. Yu、D. Parker、M. Kwiatkowska 和 L. Feng，“通过自适应保形预测实现动态智能体环境下的安全部分可观测马尔可夫决策过程在线规划”，IEEE机器人与自动化快报，2024年。[18] S. Herse、J. Vitale 和 M.-A. Williams，“信任校准的仿真证据：在人机协作中使用结合信号检测理论的部分可观测马尔可夫决策过程调整智能体特征，以优化任务结果”，《国际社会机器人学杂志》，第16卷，1403，2024年6月，第1381页。[19] Z. Zahedi、M. Verma、S. Sreedharan 和 S. Kambhampati，“信任感知规划：对迭代式人机交互中的信任演化进行建模”，载于2023年ACM/IEEE人机交互国际会议论文集，2023年，第281–289页。[20] D. Sadigh、S. Sastry、S. A. Seshia 和 A. D. Dragan，“利用对人类行为的影响进行自动驾驶汽车规划”，载于《机器人：科学与系统》，第2卷。美国密歇根州安娜堡，2016年，第1–9页。[21] A. Xu 和 G. Dudek，“Optimo：用于非对称人机协作的在线概率信任推断模型”，载于第十届ACM/IEEE人机交互国际会议论文集，2015年，第221–228页。[22] X. Hu、K. Akash、S. Mehrotra、T. Misu 和 M. Steyvers，“互惠与意识如何塑造对人工智能的亲社会行为”，载于认知科学学会年会论文集，第46卷，2024年。[23] B. Fogg 和 C. Nass，“用户如何回报计算机：一项展示行为改变的实验”，载于CHI’97计算系统中的人为因素扩展摘要，1997年，第331–332页。[24] J. W. Crandall、M. Oudah、Tennom、F. Ishowo-Oloko、S. Abdallah、J.-F. Bonnefon、M. Cebrian、A. Shariff、M. A. Goodrich 和 I. Rahwan，“与机器合作”，《自然·通讯》，第9卷第1期，第233页，2018年。[25] V. Srinivasan 和 L. Takayama，“请帮帮我：机器人向人类寻求帮助时的礼貌策略”，载于2016年CHI计算系统中的人为因素会议论文集，2016年，第4945–4955页。[26] L. E. Baum、T. Petrie、G. Soules 和 N. Weiss，“马尔可夫链概率函数统计分析中的一种最大化技术”，《数学统计年刊》，第41卷第1期，第164–171页，1970年。

本文已发表于IEEE机器人与自动化快报。此版本为作者版本。©2026 IEEE。保留所有权利，包括文本与数据挖掘以及用于训练人工智能和类似技术的权利。允许个人使用，但转载/再分发须经IEEE许可。更多信息请参阅 https://www.ieee.org/publications/rights/index.html。

[27] M. Toussaint、S. Harmeling 和 A. Storkey，“用于求解（部分可观测）马尔可夫决策过程的概率推断”，技术报告EDI-INF-RR-0934，爱丁堡大学信息学院，技术报告，2006年。[28] F. Faul、E. Erdfelder、A. Buchner 和 A.-G. Lang，“使用G*Power 3.1进行统计功效分析：相关性与回归分析检验”，《行为研究方法》，第41卷，1160，2009年4月，第1149页。[29] H. Kurniawati、D. Hsu 和 W. S. Lee，“SARSOP：通过近似最优可达信念空间实现高效的基于点的部分可观测马尔可夫决策过程规划”，载于《机器人：科学与系统》，第2008卷。Citeseer，2008年。

本文已发表于《IEEE机器人与自动化快报》。这是作者版本。©2026 IEEE。保留所有权利，包括文本与数据挖掘以及人工智能和类似技术训练相关权利。允许个人使用，但转载/再分发需获得 IEEE 许可。更多信息请参阅 https://www.ieee.org/publications/rights/index.html。