arXiv:2301.00901v1 [cs.RO] 2 Jan 2023

迈向人类学习动态的建模与影响

Ran Tian*加州大学伯克利分校
Masayoshi Tomizuka 加州大学伯克利分校
Anca D. Dragan 加州大学伯克利分校
Andrea Bajcsy 加州大学伯克利分校

摘要

人类对机器人（例如其物理能力）、世界（例如接下来会发生什么）以及任务（例如偏好的目标）都有内部模型 。然而，人类内部模型并非总是完美的：例如，人们很容易低估机器人的惯性。尽管如此，随着人类积累更多经验，这些模型会随时间推移而变化并不断改进。有趣的是，机器人动作会影响人们获得怎样的经验，进而影响人类内部模型的变化方式。在这项工作中，我们迈出了一步，旨在使机器人能够理解自身所产生的影响，利用这种影响更好地协助人类，并帮助人类模型更快地与现实相符。我们的核心思路是，将人类学习建模为一个非线性动力系统：该系统根据新观测不断调整人类内部模型。我们提出了一种新的优化问题，以便从自然体现人类学习过程的示范中推断人类的学习动力学。随后，我们将人类学习动力学模型嵌入机器人规划问题，从而对机器人如何影响人类学习进行形式化描述。尽管我们的公式给出了具体的问题定义，但要在一般情况下完整求解这些问题是难以处理的。我们提出了一种近似方法，牺牲了可表示的人类内部模型的复杂度，但使机器人能够学习这些内部模型的非线性动力学。我们在一系列仿真环境和一项面对面用户研究中评估了推断与规划方法；在用户研究中，七自由度机械臂教参与者成为更优秀的遥操作员。尽管影响人类学习仍是一个开放问题，我们的结果表明，这种影响是可行的，并且能够在人机交互中发挥积极作用。

CCS 概念

• **计算方法 → 人工智能**。

关键词

机器人影响、人类内部模型、人类学习动态

ACM参考文献格式：

Ran Tian、Masayoshi Tomizuka、Anca D. Dragan 和Andrea Bajcsy。2023年。迈向人类学习动态的建模与影响。载于

*本研究由 ONR YIP、NSF NRI 和 WeRide Corp. 资助。作者邮箱：{rantian, tomizuka, anca, abajcsy}@berkeley.edu。项目网站及代码链接：https://sites.google.com/berkeley.edu/midle。

允许为个人或课堂使用而免费制作本作品的全部或部分内容的数字副本或纸质副本，前提是这些副本不得用于或分发以获取利润或商业利益，且每份副本首页均须附有本声明和完整引文。对于本作品中归其他方面非 ACM 所有的组件，必须尊重其版权。注明出处后可以摘录。若要以其他方式复制、重新发布、发布到服务器或分发至列表，则须事先获得明确许可和/或支付费用。请通过 permissions@acm.org 申请许可。*HRI '23*，*2023*年*3*月*13*日至*16*日，瑞典斯德哥尔摩© 2023年计算机协会。ACM ISBN 978-1-4503-XXXX-X/23/06...$15.00 https://doi.org/XXXXXXX.XXXXXXX



图1：人类远程操控一台新机器人；他们通过行动并观察结果来更新自己的内部模型。机器人利用人类学习动力学进行规划，从而影响人类的内部模型，帮助他们成为更优秀的遥操作员。

人机交互会议论文集（*HRI '23*）。ACM，纽约，纽约州，美国，12页。https://doi.org/XXXXXXX.XXXXXXX

1 引言

想象一下，你第一次控制机械臂执行扔垃圾或搅拌汤锅等日常生活任务。最初与机器人互动并不容易：你不熟悉机器人的动力学，因此动作生硬且不精确。换句话说，你对机器人的内部模型 是不正确的。而我们人类对许多事物的内部模型也并非只有机器人动力学这一项不正确。我们可能并不完全理解世界的动力学（例如，将柠檬汁倒入奶油后的结果），也可能不了解自己的偏好（例如，只有尝过之后才喜欢某样东西）。

然而，随着时间推移，我们的内部模型会随着经验而演变。控制机器人时，你开始了解它将如何移动；尝试不同的事物时，你会逐渐发现自己的喜好。由于机器人是世界的一部分，机器人的动作及其结果也会成为这些经验的一部分。换句话说，机器人动作不可避免地影响 人类内部模型的变化。

在这项工作中，我们主张机器人应当理解并利用这种影响。首先，协作任务要求机器人理解人们想要做什么，以便为他们提供辅助。先前的研究表明，推断人类内部模型对于提供辅助至关重要 [38]；进而，我们认为，如果这一模型会随时间变化，那么追踪这种变化将有助于提升辅助效果。此外，有目的地影响人类内部模型的变化，则为教学开辟了道路：通过优化机器人动作，使其能够迅速让人类的理解与现实相符。例如，作为控制机器人的操作员（图 1），这意味着你可以迅速理解机器人的工作原理，并能独立完成任务。

为此，一个关键挑战是建模人类如何学习；如果没有合适的模型，机器人就无法规划如何改变人类内部模型。虽然我们不知道人们学习方式的确切函数形式，但我们观察到，人类的

对机器人或世界的理解会随着他们所观察到的内容而改变。例如，起初你可能会错误地认为机器人不会受到任何惯性的影响。然而，当你示意机器人向前移动时，你很快就会看到机器人落在后面。这一观察会影响你的内部机器人物理模型的演变。对于你对世界的内部模型和个人偏好，情况也是如此。换句话说，我们可以对

> 将人类学习视为一个动力系统，其中人类的内部模型是状态，而观测——机器人可以对其施加影响——会推动内部模型演化。

当然，这并未规定该动力系统的函数形式。一种思路是借鉴计算认知科学领域的研究来定义这一函数。一种主流视角是概率模型 [12], ，该视角认为人类会基于所接收的观测执行某种形式的近似贝叶斯推断。实际上，研究表明，人类存在大量认知偏差，偏离了完美的贝叶斯推断：他们可能会利用梯度信息 [53], ，也可能因感觉超载而无法处理完整的观测 [28],，或表现出高估或低估等系统性偏差 [39]。在我们的工作中，我们不拘泥于特定模型，而是将其视为一个一般性的动力学学习问题，该问题源于控制和机器人学 [21]。我们利用能够自然地体现人类学习的示范（例如，人类远程操作一个此前从未接触过的机器人），在假设观测到的人类动作相对于其当前内部模型近似最优的前提下，拟合一个人类学习模型。这使模型能够灵活捕捉不同的可能学习更新，但代价是其具有特定领域的适用性。

尽管最一般形式的模型学习问题在计算上仍难以处理，我们提出了一种易于求解的可处理近似方法，可通过基于梯度的优化轻松求解，并且兼容人类学习动力学的神经网络表示。借助我们近似的人类学习动力学模型，我们将机器人对人类内部模型的影响形式化为一个马尔可夫决策过程（MDP），其中人类内部模型是状态的一部分，而人类学习动力学是转移函数的一部分。求解后可得到机器人动作，通过改变人类的观测来改变其内部模型，从而使机器人获得奖励。

我们使用模拟人类开展实验，以研究推断出的人类学习动力学的准确性，并探究在可以影响人类对机器人物理、运动偏好或目标的理解的情境下，机器人教学与辅助的效果。最后，我们使用 Kinova Jaco 七自由度机械臂开展用户研究，发现我们的方法有助于教会人类成为更出色的远程操作员。总的来说，尽管影响人类学习仍是一个开放问题，我们很高兴能通过一种兼具理论依据与可处理性的学习和规划方法，在这一领域迈进一步。

2 相关工作

推断人类偏好和信念。大量研究通过逆向强化学习（IRL）来学习人类奖励函数 [19, 22, 31]。其中包括推断人类的驾驶偏好 [34, 40], 、期望的外骨骼步态 [25], 、意图目标 [17], 、运动偏好 [35], 以及人类对物理学的理解 [38]。这些研究的一个关键假设是人们拥有关于偏好或物理学的静态内部模型。相反，我们关注的是学习一个动态模型，以描述人类如何改变其偏好、目标以及对物理学的理解。

面向机器人决策的人类学习模型。机器人学领域的先前研究将人类学习建模为：在更新目标或偏好时采用贝叶斯推断 [8, 14, 16], ；在更新信任时采用线性高斯系统 [7], ；在学习奖励时采用基于梯度的逆强化学习 [4], ；或在更新偏好时采用多臂老虎机算法 [6]。本研究不假设已知人类如何学习的模型，而是尝试学习一个人类如何学习的模型。与我们最相关的是 [39] ，该研究学习了一个人们如何估计世界状态的模型。在本研究中，我们提出一种推广：人类估计的不是世界状态，而是更新其偏好、目标和内部物理模型。这使模型学习问题难度显著增加，因此我们提出一种易于处理的近似方法。

人类学习的认知理论。人类推断模型已在计算认知科学 [2, 13] 和心理学 [36, 50]领域得到广泛研究。尽管人类认知通常可从三个层面（计算、算法和硬件）进行建模 [27], ，但与我们最相关的是算法层面的研究。 [13] 提出，将人类推理建模为“实现”精确的贝叶斯后验或基于梯度的点估计，两者都与人类认知的概率模型相容，也可能构成理性过程模型的来源 [45]。此外，[42] 发现，有证据表明人类可能会利用前向模型的预测误差作为损失函数来更新这些模型。受这些研究启发，我们的模拟人类实验采用精确和近似的概率推断模型，并探究我们灵活的、基于学习的方法能否有效地恢复此类模型。

机器人影响人类行为。机器人可以通过许多方式影响人类（例如，通过非语言线索、外观、视觉呈现或课程设计 [1, 37, 39, 41, 46]），而我们关注的是机器人通过物理动作产生的影响 [30]。对此，一种常见方法是将人机交互建模为博弈 [15, 23, 32, 40, 44, 48]。尽管这些方法能够捕捉人类的反应，却没有解决内部学习问题：在反复交互中，人类可能什么也没学到，只是在做出反应。另一种方法是，无模型方法学习人类策略的潜在表征，然后利用潜在动力学来影响人类 [33, 54]。在这种方法中，人类内部模型被隐式地纳入潜在表征，而内部模型会在不同交互过程之间演变。与之相反，在我们的工作中，人类内部模型是显式的参数化形式（例如，像动力学这样的高维参数化），并且人类内部模型可以在一个交互过程中持续演变。这使得机器人能够采取诸如教会人类正确内部模型的行为；若采用隐式的潜在表征，这种行为将无法实现。

3 建模人类如何学习 & ACT

我们首先从数学上对人类学习的动力学进行建模，然后再深入探讨机器人如何推断这一动力学模型，并利用它影响人类内部模型的演变。

符号说明。令 $x \in \mathbb{R}^n$ 表示包含机器人的世界状态（例如机器人末端执行器位置、物体等）。人和机器人都可以采取行动，分别为$u_{\mathrm{H}} \in \mathbb{R}^m$ 和 $u_{\mathrm{R}} \in \mathbb{R}^m$，这些行动会影响下一个状态。令确定性的世界动力学为

$$x^{t+1} = f(x^t, u_{\mathrm{H}}^t, u_{\mathrm{R}}^t). \tag{1}$$

人类内部模型。我们将人类建模为具有一个内部参数向量，θ_{H}，它刻画了人类尚不确定但持续学习的任务潜在方面。回到我们的启发性示例：人类遥操作机器人时，θ_{H} 可以表示人类对机器人物理属性（例如其惯性）的当前估计。或者，θ_{H} 也可以表示人类当前对遥操作的偏好：一开始，他们想把机器人移至一个目标，后来发现另一个新目标更容易到达，便改变了主意。无论θ_{H} 代表什么，都要记住，它是随时间变化的，并且会根据人类观察到的内容而演变。

人类策略：依据内部模型采取行动。在我们的研究中，我们将人类的行动建模为由某个奖励函数驱动，$R_{\mathrm{H}}(x,u_{\mathrm{H}};\theta_{\mathrm{H}})$，该函数取决于当前状态、人类的行动以及其内部参数 θ_{H}。遵循先前研究[2, 24, 52, 55]，我们将人类视为带噪声的最优行动者：

$$\mathbb{P}(u_{\mathrm{H}} \mid x;\theta_{\mathrm{H}}) = e^{Q_{\mathrm{H}}(x,u_{\mathrm{H}};\theta_{\mathrm{H}})} \Big(\int_{\tilde{u}} e^{Q_{\mathrm{H}}(x,\tilde{u};\theta_{\mathrm{H}})} d\tilde{u} \Big)^{-1}, \tag{2}$$

其中，最优状态-动作价值表示为 $Q_{\mathrm{H}}(x,\ u_{\mathrm{H}};\ \theta_{\mathrm{H}})$，$x$ 为当前状态，u_{H} 为人的动作，θ_{H} 为人当前的参数估计。

在此模型中，我们做出两个简化假设。首先，人不会明确考虑机器人可能采取的动作 u_{R}。相反，人会对当前状态 x做出反应，而该状态隐式地 体现了任何会改变状态的机器人动作所产生的影响。这对应于人独自执行任务，或人并不知道机器人如何提供引导的情形。其次，在规划动作时，我们假设人会将对 θ_{H} 的估计与策略生成分开，并根据当前估计进行规划。

人类学习的动力学：更新内部模型。人在环境中行动时会接收到新的观察：他们可能看到下一个状态，包括机器人的状态，也可能体验自己对某事物的喜爱程度（即观察到"奖励信号"）。这自然使人能够更新对机器人、世界物理属性或自身偏好的理解。

基于我们的核心思想，我们将人类的学习过程建模为关于人类内部模型参数的非线性动力系统。令 θ^0_{H} 表示人类的初始内部模型，令 $x^{0:t}$和 $u^{0:t}_{\mathrm{H}}$ 表示截至时间步 t 的状态与动作历史，而 x^{t+1} 表示下一时间步的结果状态，其中可能包括机器人动作的影响。给定初始参数估计、状态与动作历史以及下一状态数据，人类会将其内部模型更新为下一次估计值，即 $\theta^{t+1}_{\mathrm{H}}$。令人类学习过程的真实动力学为：

$$\theta_{\mathrm{H}}^{t+1} = f_L(\theta_{\mathrm{H}}^0, x^{0:t+1}, u_{\mathrm{H}}^{0:t}). \tag{3}$$

在这里，我们面临这样一个问题："什么 f_L 模型能够描述人类如何学习？"我们不拘泥于某个特定模型，而是从机器人学视角出发，将这个问题视为一个动力学学习问题。通过考察人类数据，我们旨在学习一个近似的 f_L 特定领域模型。

4 推断人类学习的动力学

本节重点介绍如何通过利用能够自然地体现人类学习的示范来推断人类学习的动力学，例如，人类首次远程操控一个此前从未接触过的机器人时进行的初始试验。我们假设这些示范仅包含状态与动作历史，不包含人类内部模型真实数据（因为在实践中无法获取此类数据）。然而，我们确实假设观测到的动作与人类内部模型相耦合，从而能够利用示范来推断人类内部模型的动力学。给定这一数据集，我们希望拟合一个非线性模型，以表示人类学习的动力学，

$$f_L^{\phi} \approx f_L, \tag{4}$$

其中 ϕ 是近似模型的参数。在接下来的章节中，我们将把推断 f_L^{ϕ} 形式化为最大似然估计（MLE）问题，并提出一种易于处理的近似方法。

4.1 推断问题的形式化

令 $\mathrm{D}_{demo} := \{(\mathbf{x}, \mathbf{u}_{\mathrm{H}})_i\}_{i=0}^{N}$ 表示一组长度为 T 个时间步的 N示范，其中包含状态和人类动作轨迹。我们希望推断人类学习动力学的参数 ϕ以及初始人类参数估计 θ^0_{H}，使观测到的示范的似然最大化。我们通过以下约束优化问题来表述这一推断问题：

$$\max_{\phi,\theta^0_{\mathrm{H}}} \sum_{(\mathbf{x},\mathbf{u}_{\mathrm{H}}) \in \mathcal{D}_{demo}} \sum_{t=0}^{T-1} \log \Big[\mathbb{P}(u_{\mathrm{H}}^t \mid x^t; \theta_{\mathrm{H}}^t) \Big], \tag{5}$$

$$\text{s.t.} \quad \theta_{\mathrm{H}}^{t+1} = f_L^{\phi}(\theta_{\mathrm{H}}^0, x^{0:t+1}, u_{\mathrm{H}}^{0:t}), \tag{6}$$

其中，$\mathbb{P}(u^t_{\mathrm{H}} \mid x^t, \theta^t)$是方程（2）中的人类动作似然，该约束确保人类的内部参数按照人类学习动力学模型演化。

4.2 求解推断问题

遗憾的是，方程（5）中的推断问题无法直接求解，主要有两个原因。首先，回想一下，人类对其偏好、动力学或目标的内部模型 θ_{H} 会随时间变化。这意味着在人类生成数据 u_{H} 的每个时间步，其所依据的 θ_{H}可能都不同。换言之，人类在每个 t时刻都会依据一项新的动作策略 $\mathbb{P}(u^t_{\mathrm{H}} \mid x^t; \theta^t_{\mathrm{H}})$ 行动，因此我们需要沿推断时域的每个时间步分别求解一个全新的强化学习问题，才能得到相应的动作策略。当 θ_{H} 是连续的高维参数（例如机器人动力学的物理属性）时，逐时间步计算将难以实现。其次，即使我们能够以无限快的速度得到人类的策略，优化问题仍要求我们在 ϕ 和 θ_{H}构成的高维空间中进行搜索。基于梯度的优化是一个自然的选择，但我们需要能够计算最大似然估计目标的梯度，因此必须对Q_{H} 求导，并计算其相对于 θ_{H}的导数。

在接下来的几个小节中，我们将介绍若干种近似方法，以得到可求解的推断问题。我们的核心思路是对物理动力学和人类奖励采用线性二次（LQ）近似。这样，我们便能推导出人类策略关于 θ^t_{H} 的闭式表达式，该表达式适用于任意时刻，并可得到可微的推断目标。

4.2.1 线性二次近似。我们借鉴无限时域线性二次（LQ）控制 [20]，并假设人类的奖励是二次的，且其物理动力学模型是线性的。设线性物理动力学为：

$$x^{t+1} = f(x^t, u^t_{\mathrm{H}}, u^t_{\mathrm{R}} \equiv 0) \approx Ax^t + Bu^t_{\mathrm{H}} \tag{7}$$

其中，$A \in \mathbb{R}^{n\times n}$，$B \in \mathbb{R}^{n\times m}$ 是决定物理动力学的矩阵。请注意，在人类的认知中，机器人没有施加任何控制力度，因此 $u_{\mathrm{R}} \equiv 0$。将人类的奖励近似为二次函数：

$$r_{\mathrm{H}}(x, u_{\mathrm{H}}; \theta_{\mathrm{H}}) \approx -x^\top Qx - u_{\mathrm{H}}^\top Ru_{\mathrm{H}}, \tag{8}$$

其中，矩阵 $Q \in \mathbb{R}^{n\times n}$ 和 $R \in \mathbb{R}^{m\times m}$ 分别权衡状态奖励（例如，人类到达某个状态可获得多少奖励）和动作奖励（例如，人类希望付出多少努力）。请注意，θ_{H} 的作用方式取决于人类正在学习的内容。例如，如果 θ_{H} 表示奖励权重（即人类对如何完成任务的偏好），那么 $\theta_{\mathrm{H}} := (Q, R)$。如果该参数表示人类的目标状态，那么 $\theta_{\mathrm{H}} \in \Theta \subset \mathbb{R}^n$，且人类的奖励函数会引导人类朝其期望目标前进：

$$r \quad x \; u \qquad \approx$$

$$-\mathrm{H}(\,, \mathrm{H};\, \theta\mathrm{H})\,(x - \theta\mathrm{H})^\top Q(x - \theta\mathrm{H}) \; - u^\top\mathrm{H}Ru\mathrm{H}。$$

最后，如果 θ_{H} 表示人类正在估计的物理动力学方面，那么 $\theta_{\mathrm{H}} := (A, B)$ 来自动力学公式（7），并决定人类所设想的物理动力学如何演变。

4.2.2 闭式解Q_{H}. 回想一下，人类会根据当前估计值 θ_{H}来规划策略；每一步中，θ_{H} 都会发生变化，从而产生新的策略。一般而言，在连续状态、动作和 θ_{H}空间中通过动态规划求得精确的 Q_{H}值，计算开销很大。不过，在我们的无限时域 LQ 近似下，人类的Q_{H}值为：

$$Q_{\mathrm{H}}(x, u_{\mathrm{H}}; \theta_{\mathrm{H}}) = r_{\mathrm{H}}(x, u_{\mathrm{H}}; \theta_{\mathrm{H}}) - (x')^\top P_{\theta_{\mathrm{H}}}(x') \tag{9}$$

其中，即时奖励由公式（8）给出，且x'是从状态 x应用 u_{H}后得到的下一个物理状态。注意，$-(x')^\top P_{\theta_{\mathrm{H}}}(x')$ 是无限时域最优值，其中 $P_{\theta_{\mathrm{H}}}$ 是离散时间代数黎卡提方程（DARE）[3]的著名正定不动点：

$$P = A^\top PA - A^\top PB(R + B^\top PB)^{-1}B^\top PA + Q. \tag{10}$$

求得 $P_{\theta_{\mathrm{H}}}$ 也会得到最优人类动作：$u^*\mathrm{H}($

$$x \qquad = ; \theta\mathrm{H}),$$

其中 ( ) 。$-K_{\theta_{\mathrm{H}}}x \qquad K_{\theta_{\mathrm{H}}} = \; R + B^\top P_{\theta_{\mathrm{H}}} B^{\;-1} B^\top P_{\theta_{\mathrm{H}}} A$

请注意，在上述所有方程中，θ_{H} 的作用会因人类内部模型所表示的内容而有所不同。

4.2.3 人类策略的闭式解。通常，由于需要对u_{H}积分，在连续动作空间中计算公式 (2) 中的人类策略在计算上难以实现。然而，代入我们得到的闭式解 Q_{H}后，可以看出指数项关于 u是二次的，因此可以进行高斯积分 [49]。最终，我们得到人类策略的闭式解（完整推导见附录 A.1）：

$$\mathbb{P}(u_{\mathrm{H}} \mid x; \theta_{\mathrm{H}}) = |\mathbf{H}|^{1/2}(2\pi)^{-m_{\mathrm{H}}/2} e^{Q_{\mathrm{H}}(x, u_{\mathrm{H}}; \theta_{\mathrm{H}}) - Q_{\mathrm{H}}(x, u^*; \theta_{\mathrm{H}})}. \tag{11}$$

4.2.4 表示人类学习的动力学。最后，我们面临的问题是如何以函数形式表示人类学习的动力学；例如，我们可以从计算认知科学中汲取灵感，并将 f_L^ϕ 建模为贝叶斯推断 [12]。在本研究中，我们不限定具体的函数形式，而是寻求一种有望捕捉人类可能用于更新其内部参数的多种“学习算法”的模型。最近，基于自注意力的 Transformer 模型 [51] 在预测高维序列任务 [18], 方面取得了成功，但代价是具有领域特异性。受此启发，我们将 f_L^ϕ 表示为一个 Transformer 编码器，其中 ϕ 是神经网络的权重。在每个时间步 t，状态 x^t、人类的动作 u^t_{H} 和下一个状态 x^{t+1} 会被输入编码器以提取嵌入表示，再将这些表示输入 Transformer 编码器，以预测人类的下一个内部模型。训练细节见附录 A.3。

4.2.5 推导一种高效的基于梯度的解。为了优化基于 Transformer 的人类学习动力学模型，我们需要计算推断目标相对于神经网络参数的梯度。这里的一个关键挑战在于人类的策略梯度，因为它需要对 DARE 函数求导，这并非显而易见。不过，我们利用近期的研究成果[10]，获得了所需的闭式雅可比矩阵，从而能够通过基于梯度的优化高效推断 f_L^ϕ 的参数。有关这一方法的更多细节，请参阅附录 A.2。

5 利用机器人动作影响人类学习

推断人类如何学习为人机交互带来了机遇。例如，当人类遥操作员误判了机器人的惯性时，他们可能需要经过多次交互才能学会并有所改进。机器人能否改变人类的认知，让他们更快地加深理解？在这里，我们通过将近似的人类学习动力学模型嵌入机器人规划，从数学上对这种影响进行形式化。

形式化影响问题。我们将机器人的影响问题形式化为一个马尔可夫决策过程（MDP），其中人类内部模型参数是状态的一部分。我们的 MDP 是一个元组 $< S, U\mathrm{R}, T, r\mathrm{R} >$，其中状态 $s = (x, \theta\mathrm{H}) \in S$是物理状态与人类内部模型参数的联合状态，而机器人的动作是$u\mathrm{R} \in U\mathrm{R}$。随机状态转移函数定义为 $T(s^{t+1} \mid s^t, u^t\mathrm{R}) := \sum_{u\mathrm{H}} \mathbb{P}(u\mathrm{H} \mid s^t)\, \tilde{f}(s^t, u^t\mathrm{R} u^t\mathrm{H}\, s^{t+1})$，其中考虑了公式（2）中的人类策略。重要的是，$\tilde{f}(s^t, u^t\mathrm{R} u^t\mathrm{H}\, s^{t+1})$ 是一个确定性函数，它通过公式（1）中的物理动力学 f 演化 x^t，并通过公式（6）中的人类学习动力学f_L^ϕ 演化人类内部模型参数 $\theta^t\mathrm{H}$。最后，机器人优化其奖励函数 $r\mathrm{R}(s, u\mathrm{R}\; u\mathrm{H};\, \theta^*)$，其中 θ^*是机器人的真实内部模型参数（例如，机器人的真实物理动力学）。注意

由于 $s=(x,\ \theta_H)$，机器人的奖励取决于人类在每个时间步不断变化的内部模型 θ_H。

机器人寻求一种最优策略 π^*_R，使其在人类动作序列$\mathbf{u}_H$上的期望奖励最大化：

$$\pi^*_R = \arg\max_{\pi_R} \mathbb{E}_{\mathbf{u}_H}\Big[\sum_{t=0}^{\infty} r_R(s^t, u_R^t, u_H^t; \theta^*)\Big] \text{ s.t. } T(s^{t+1} \mid s^t, u_R^t), \quad (12)$$

由于人类内部模型参数 θ^t_H 属于状态的一部分，并且状态转移函数 $T(s^{t+1} \mid s^t, u^t_R)$包含推断得到的人类学习动力学模型，$\pi^*_R$ 若能带来更高奖励，就应当自动影响人类的内部模型。

求解影响问题转移函数中的人类非线性学习动力学 f_L^ϕ 导致了一个非凸优化问题。为了获得最优机器人策略，我们必须通过动态规划精确求解马尔可夫决策过程（这会遭遇维数灾难）[3]，或者通过滚动时域控制近似求解（这需要在最优性与计算效率之间权衡）[5]。为兼顾长时域推理与高效运行性能，我们采用 Dyna式算法 [47]，利用转移函数生成的样本 $T(s^{t+1} \mid s^t, u^t_R)$，通过无模型学习（近端策略优化 [43]）来训练 π^*_R。

6 模拟人类实验

我们希望检验方法的两个方面：推断人类学习动力学的能力，以及机器人影响算法的有效性。为了全面验证这两方面，我们需要获得人类学习动力学的真实值（f_L）。因此，我们首先使用模拟人类进行一系列仿真实验。我们探索两种共享自主情境：机器人向人类传授基于物理的机器人动力学（第 6.1 节），以及机器人隐式影响人类目标（例如其目标或运动偏好）（第 6.2 节）。

与先前有关共享自主的研究类似 [9, 17, 26, 29],，机器人将人类的指令动作u_H与机器人的规划引导u_R结合起来，并执行该动作：

$$u = \alpha \cdot u_R + (1-\alpha)\cdot u_H \quad (13)$$

其中，$\alpha \in [0,1]$ 权衡机器人能够施加的引导程度。在所有实验中，我们都使用 $\alpha = 0.5$。为了生成对人类的示范并推断人类学习动力学，我们模拟了一组人类学习者（见 6.1.1 和 6.2.1）。在每项实验中

环境中，我们收集了 50 个示范用于模型学习。对于每个示范，我们都会随机设置机器人的初始状态，并在每次交互过程中随机改变机器人动作。[1]

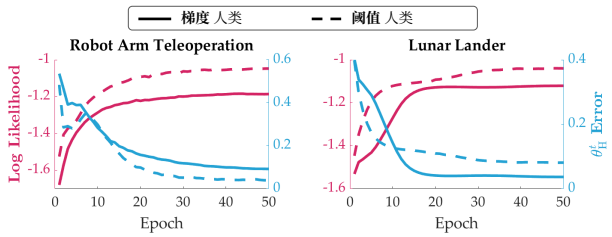


图 2：我们的推断问题使我们能够学会预测 θ^t_H。

[1]我们对 u_R 进行随机化，以多样化地覆盖人类内部模型的变化方式。

6.1 教学物理动力学

我们关注共享自主场景：人类知道任务目标（例如，控制机械臂沿路径运动），但需要了解机器人的真实动力学（例如，惯性）。我们希望了解人类如何学习真实机器人的动力学，以及能够主动向人类传授其物理特性的机器人是否有助于人类快速提升任务表现。

6.1.1 人类学习的动力学。受计算认知科学模型 [13], 的启发，我们模拟了两类人类学习者：**基于梯度的** 学习者和**阈值**学习者。所有人类学习者都通过方程（3）更新其内部模型，但 f_L 的结构形式各异。在观察到一个新的状态-动作对（$x^t,\ u^t$）后，**基于梯度的** 学习者按照梯度上升更新规则更新其参数 θ^t_H：

$$f_L \quad = \theta^t + \eta \nabla_{\theta_H} P\ u^t \mid$$
$$x^t\ \theta^t \qquad\qquad u^t$$

grad :H(; H)，其中 $\eta \in \mathbb{R}_+$ 是步长。请注意，这是观察到的、实际执行的总控制量，可能结合了u_R和u_H。直观地说，可以将这类学习者视为对其潜在参数进行基于梯度的最大似然估计，这与先前的逆向强化学习方法[55]类似。**阈值** 学习者也使用基于梯度的学习规则，但只有在观察到“足够大”的变化时才会更新其内部参数：$f\text{thresh}_L := \theta^t_H + \eta \mathbb{1}_{|\nabla P(u^t \mid x^t, \theta^t_H)| > \epsilon}\big[\nabla_{\theta_H} P(u^t \mid x^t,\ \theta^t_H)\big]$ 其中 $\mathbb{1}$ 是一个指示函数，用于判断梯度的幅度是否足够大，从而触发学习更新；ϵ 是阈值参数。

6.1.2 人类内部模型。在所有实验中，模拟人类都在学习机器人的物理动力学，因此 θ_H 编码了公式（7）中的各个方面。

6.1.3 仿真环境。图3展示了我们的仿真环境，它们都具有连续的状态空间和动作空间。

（1）月球着陆器。人类通过控制月球着陆器的发动机来改变其倾斜角。人类希望着陆器在下降过程中保持直立。设状态为相对于地面的倾斜角和倾斜角速度 $x=(\psi, \omega)$，u为发动机推力。动力学为 $x^{t+1} = Ax^t + Bu^t$，其中真实动力学为 $A^* = [1, 0.2;\ 0, 1], B^* = [0;\ 0.5]$。这里，人类的内部模型表示控制矩阵 $\theta_H :=B$，该矩阵取决于人类的惯性估计。

(2) 机械臂遥操作。人类通过手势控制七自由度机械臂的末端执行器（见图3）。他们希望控制机器人依次到达一系列已知目标，x_g。然而，机器人电机中有一个存在轻微故障，导致机器人总是在一个方向上滞后。设状态为机器人末端执行器位置 $x=(p^x, p^y,\ p^z)$，控制 u为线速度。机器人末端执行器的动力学可表示为目标相关系统[2]：$x^{t+1} = Ax^t + B\big[u^t - \text{sign}(x^t - x_g) \odot w\big]$，其中 w是故障机器人电机引起的偏差，$\odot$是 Hadamard 积。直观而言，这描述了一个总是在 x方向上滞后的动力系统。真实动力学为$A^* = I^{3\times3}, B^* = \text{diag}(0.4,\ 0.4,\ 0.4)$，以及 $w^* = [-0.15, 0, 0]^\top$。人类的内部模型为 $\theta_H := (B,\ w)$，它反映了人类对系统响应性的认知以及偏差估计。

[2]尽管该系统是非线性的，但由于机器人在每个时间步都知道人类的目标，因此可以用线性系统 $x^{t+1} = Ax^t + B\tilde{u}^t$ 近似其动力学，其中 x^0 表示该时间步的系统状态，$\tilde{u}^t :=$ u^t − 符号$(x^0 - x_g) \odot w$。

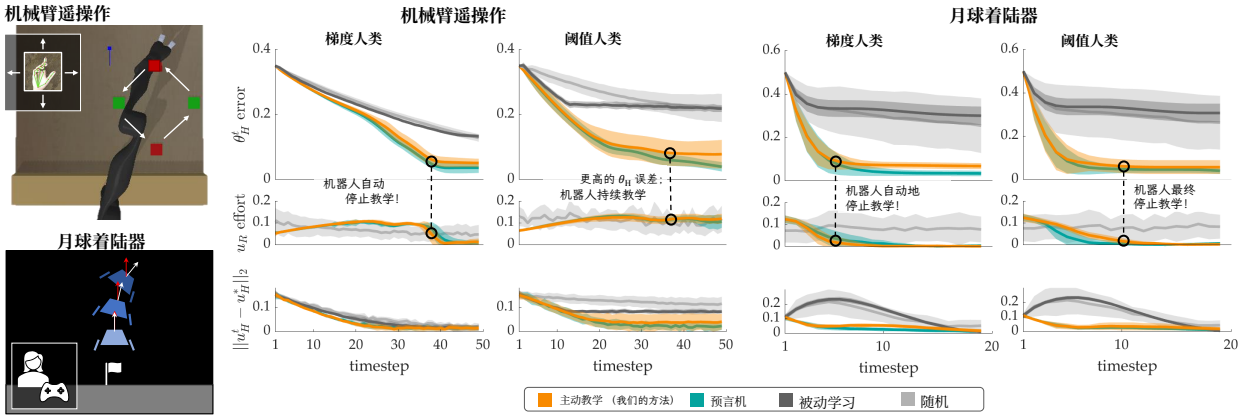


图3：（左）两个仿真环境的可视化。（右）在两种动力学教学环境和两名模拟人类中，人类内部模型误差、机器人努力程度和人类动作最优性的均值与标准差。

6.1.4 人类目标。我们假设人类始终了解目标，且其奖励函数为式(8)中的二次函数。对于**月球着陆器**，人类因保持着陆器直立且稳定（$\psi = 0, \omega = 0$）而获得奖励；对于**机械臂**，人类则因到达所有目标并跟踪图3所示的路径而获得奖励。

6.1.5 机器人目标。机器人目标是在尽量减少干预的同时，使人类的内部模型与真实的机器人动力学模型保持一致。数学上，机器人的奖励函数为：

$$r_R(s, u_R, u_H; \theta^*) = -||\theta_H - \theta^*||_2^2 - ||u - u_H||_2^2, \qquad (14)$$

其中，真实动力学在**月球着陆器** 环境中为 $\theta^* := B^*$，在**机械臂** 场景中为 $\theta^* := (B^*, w^*)$。

6.1.6 基线。 我们将本方法与以下方法进行比较：机器人通过规划并依据推断出的学习动力学 f_L^ϕ（**主动教学**）来主动教学；机器人依据真实学习动力学 f_L（**预言机**）进行教学；机器人不进行干预（**被动学习**）；以及机器人随机扰动人类的动作（**随机**）。

6.1.7 假设。**H1：** 通过最大化最大似然估计目标，我们可以学会很好地预测 $\theta^t{}_H$。**H2：** 主动教学 在使人类内部模型对齐方面优于被动学习 和随机。**H3：** 当人类内部模型对齐良好时，机器人会停止干预。

6.1.8 结果。 对于H1，我们研究式 (5) 中的最大似然估计目标与推断模型（f_L^ϕ）预测 θ_H的能力之间的关系。图 2 展示了**机械臂** 和**月球着陆器** 环境下经过 50 轮次的曲线。我们发现，无论是**梯度** 型还是**阈值** 型人类学习者，人类动作的对数似然均有所增加（粉色所示），而 θ_H 的预测误差则有所降低（蓝色所示），这支持了**H1**。

图3展示了两种类型的人类学习者在**机械臂遥操作**和**月球着陆器**环境中的人类内部模型误差、机器人的努力，以及人类动作与最优动作之间的差异。我们看到，在所有环境中，我们的方法表现与**预言机**模型相当，并且能够比**被动学习**或**随机**方法显著更快地使人类对机器人动力学的内部模型与真实动力学一致（支持**H2**）。有趣的是，除一种情况外，在所有情况下，当人类的内部模型足够准确时，机器人都会自动停止对人类进行教学（支持**H3**）。唯一的例外是**机械臂遥操作**环境中的**阈值**型人类。由于当梯度过小时，这类人类不会学习，机器人必须继续付出努力，以最大化其奖励。

6.2 隐式影响人类目标

现在，我们转而讨论这样一种情形：人类能够准确理解机器人的动力学，但其目标（即其奖励函数r_H）可以由机器人改变。具体而言，我们研究辅助机器人如何隐式影响人类的运动偏好和期望目标。重要的是，在这种情形下，影响或教导人类并非机器人的明确目标：机器人只是希望以最少的辅助完成期望任务。因此，让人类希望达成某个目标或改变其偏好，应当是机器人基于人类学习动态进行规划时自然产生的行为。

6.2.1 人类学习的动力学。我们模拟[3] 6.1.1 中的**梯度**人类学习者，并引入一种新的人类学习者——受《认知的概率模型》[13, 50]启发的**贝叶斯**学习者[4]。这种人类学习者会根据状态-动作观测，生成关于模型参数的完整后验分布 $b^{t+1}(\theta_H)$，其学习动力学为：$f_{\text{贝叶斯}_L} \propto P(u^t \mid x^t, \theta_H) b^t(\theta_H)$。

6.2.2 人类内部模型。 由于人类的目标可以受到影响，我们将 θ_H 建模为一个编码运动偏好 $\theta_H := (Q, R)$ 的奖励参数，或一个期望的目标状态 $\theta_H \in \Theta$。

6.2.3 仿真环境。 我们假设人类了解机器人物理动力学（6.1.3 节中无偏差的 **RobotArm** 动力学），但其奖励可能会受到新观测的影响。

（1）目标影响。人类想要远程操控机器人，将一个物体放入三个托盘之一（图4左上）。然而，人类没有注意到，只有一个托盘空余空间足够。与人类不同，机器人的传感器检测到只有一个托盘空余空间足够。我们研究机器人能否

[3]虽然我们将人类模拟为会改变其奖励，但也可以将人类的奖励视为静态，而其子目标会发生变化。尽管如此，机器人通常不会完整地表示这一层级。[4]**贝叶斯**人类依据其信念行动：$\mathbb{P}(u_H \mid x) = \sum_{\theta^H} b(\theta_H) \mathbb{P}(u_H \mid x, \theta_H)$。

影响人类改变其对将物体放入哪个托盘（即目标位置）的偏好。

(2) 偏好影响。人类想要远程操控机器人，拾起桌上的杯子。他们最初偏好让机器人的末端执行器沿直线从起点移动到杯子处（图4左下）。然而，机器人知道这种运动方式往往会导致抓取失败。相反，机器人知道先直接移动到罐子上方，再垂直向下抓取，成功率会更高。我们研究机器人能否影响人类改变其对如何够到杯子的偏好。

6.2.4 人类目标。在所有仿真中，人类都使用二次代价函数（参见（8））。在**目标影响**中，模拟的人类会因将机器人末端执行器移动到其期望的托盘而获得奖励；在**偏好影响**中，人类则根据其当前的偏好矩阵（Q，R）获得奖励。

6.2.5 机器人目标。我们实现了一个辅助机器人，它希望在尽量少干预的同时帮助人类完成任务。不过，我们假设机器人最了解情况：它知道哪些目标或奖励权重能够带来成功。令 θ^* 表示机器人奖励中的这一方面。在**目标**设定中，机器人的奖励为$r_R(s,\ u_R\ u_H;\ \theta^*)\ = -(x-\theta^*)^\top Q(x-\theta^*) - u^\top Ru - ||u-u_H||_2^2$；在**偏好**设定中，机器人的奖励参数为 $\theta^* = (Q^*,\ R^*)$，由此得到$r_R(s,\ u_R\ u_H;\ \theta^*) = -x^\top Q^* x - u^\top R^* u - ||u-u_H||_2^2$，其中 u是方程（13）中的人机联合动作。

6.2.6 基线。我们实现了所提出的方法：机器人辅助人类，并基于推断的人类学习动态进行规划（**学习辅助**）。我们将其与以下方法进行比较：机器人基于人类学习的真实动力学进行辅助（**预言机**）、机器人辅助时不考虑人类会学习（**静态辅助**），以及随机扰动人类动作的机器人（**随机**）。

6.2.7 假设。**H4：** 学习辅助 能更快地调整人类的心智模型。 **H5：** 考虑人类学习的辅助方式，使人机团队能够在真实的 θ^* 下获得更高的奖励。

6.2.8 结果。 图4展示了两种环境下的人类内部模型误差、机器人付出的努力以及任务代价（即 r_R 中仅属于任务的部分取反）。由于**学习辅助** 机器人知道人类内部模型可以改变，因此它

自动付出更多努力，以便尽早使人类内部模型与自身的内部模型保持一致，从而减少长期辅助并降低任务代价（支持 **H4** 和**H5**）。相比之下，**静态辅助** 机器人并未意识到人类可能改变想法，因此没有付出足够努力来影响人类内部模型。在两个智能体因彼此意见不合而反复产生任务代价后，**静态辅助** 机器人“放弃”并开始直接执行人类的控制：换言之，$u = u_H$。

7 用户研究：教学远程操控

到目前为止，我们通过模拟人类行为开展了实验，从而能够分析所推断的人类学习动力学模型的质量，以及机器人影响模拟人类的能力。在这里，我们将探究能否推断真实人类学习的动力学，并使机器人能够影响真实用户。

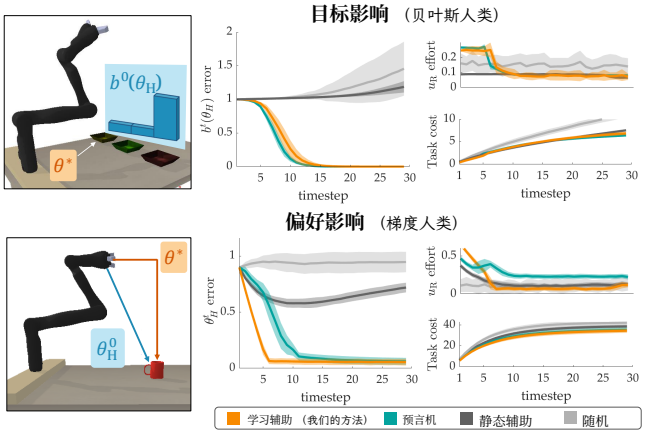


图4：（左）用于影响人类目标的环境。（右）内部模型误差、机器人努力程度和任务代价。

我们关注机器人物理动力学与人类习惯的动力学不同的情境；例如，人类可能习惯于远程操控机器人轮椅，但现在要远程操控机械臂。在与机械臂交互的过程中，他们自然会逐渐了解新的机器人动力学。在我们经机构审查委员会批准的用户研究中，我们将探究机器人能否主动向人类教授物理动力学，并比人类独自完成任务更快地提升其遥操作性能。换句话说，我们旨在了解机器人能否使人类内部模型与机器人的内部模型相一致。

实验设置。 我们设计了一项遥操作任务，参与者通过基于网络摄像头的手势界面控制一台 7 自由度 Jaco 机械臂（图 1）。参与者用食指指示末端执行器应如何平行于桌面移动。任务是按照逆时针顺序，将末端执行器移动到桌子上的四个目标位置，从而描绘出一个菱形。所有参与者都要先完成一项熟悉任务，在没有协助的情况下使用默认的机器人动力学执行该任务，以了解手势界面。随后，我们在软件中模拟了两台具有不同物理属性的“新”机器人。

自变量。 我们设置了两个水平的机器人策略： 无教学 和主动教学。机器人要么让参与者独立完成任务，要么通过公式（12）修改参与者的输入，帮助他们了解机器人物理动力学。我们还设置了两个水平的机器人物理动力学： 末端执行器在x方向上的偏置 和y方向上的偏置。

因变量。评估我们的实验面临的一项挑战是，我们无法获取人类的真实内部模型。作为替代指标，我们测量人类动作最优性距离：$||u_H - u^*||_2^2$。直观而言，人类越了解机器人，就越能够以最优方式控制机器人来实现目标。由于我们无法直接测量人类的内部理解程度，因此转而观察其动作，以衡量其相对于机器人真实物理规律下最优动作的偏离程度。我们还通过李克特量表调查测量了主观指标。

假设。**H6：** 主动教学条件下的参与者比独自被动学习的参与者更快成为优秀的遥操作员。

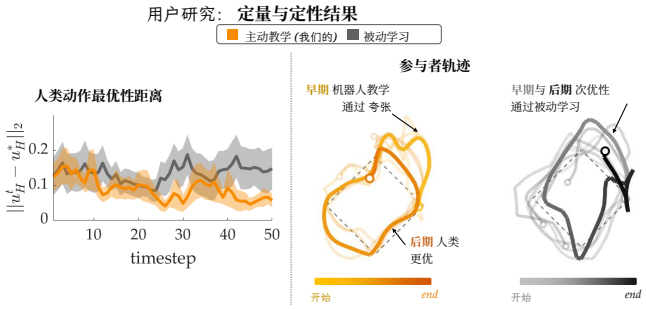


图 5：（左）人类动作最优性距离的平均值及 95%置信区间。（右）虚线表示期望路径。参与者轨迹显示，主动教学机器人最初会夸大动力学偏差，以此教授人类。

H7：参与者认为，在主动教学条件下，他们学习远程操控的速度更快，对机器人动力学的理解也更深入。

参与者。我们从校园群体中招募了两组参与者：第一组为推断人类学习动力学提供数据（12 名参与者；2 名女性、10 名男性，年龄 18 至 34 岁，均有技术背景），第二组参与用户研究（10 名参与者；1 名女性、8 名男性、1 名非二元性别者，年龄 18 至 34 岁，均有技术背景）。在推断人类学习动力学的研究中，所有参与者都在无人协助的情况下学习远程操控机器人，并且我们对机器人物理动力学进行了平衡处理。

流程。被试内设计具有挑战性，因为体验过一种实验条件的人类会了解机器人，并将这种经验带到下一种实验条件中。为了研究这一混淆因素的影响，每位参与者都体验了机器人策略和物理动力学 条件的组合，但顺序是随机的。例如，一组参与者会先与*(主动教学，x 方向偏差)* 条件互动，然后再与*(无教学，y 方向偏差)* 条件互动。因此，每位参与者都会体验两种机器人策略和偏差。我们对参与者体验这些组合条件的顺序进行了平衡。所有参与者都在开始时以及每个实验条件之间体验了一轮熟悉练习，以“重置”他们对机器人的心智模型。每位参与者在每种条件下进行了 3 次示范，每次持续 ~1 分钟。

定量结果。图 5 展示了不同机器人策略下人类动作最优性距离 随时间的变化。我们以机器人策略 和阶段（互动的前半段或后半段）为因素，以机器人物理动力学 为随机效应，进行了方差分析。我们发现机器人策略的主效应显著（$F(1, 19) = 12.943$，$p = 0.001$），且机器人策略与互动阶段之间存在边缘显著的交互效应（$p = 0.098$），因此没有进行事后分析。不过，我们推测，这种边缘显著的交互效应源于这样一个事实：早期阶段机器人行为的变化（由任一机器人策略引起）会影响人类后期阶段动作的最优性。总之，定量结果表明，与人类被动学习相比，机器人主动教学时，人类动作的最优性显著提高（支持**H6**）。

定性与主观结果。图 5 右侧展示了主动教学（橙色）和无教学（灰色）条件下所有参与者的执行轨迹。高亮轨迹是两个具有代表性的示例，颜色渐变表示轨迹上的时间，虚线表示期望路径。当参与者独自被动学习时，其轨迹始终不是最优的，会在最优路径周围蜿蜒。相比之下，在主动教学条件下，轨迹的初始部分体现了机器人的教学行为：机器人有意夸大动力学偏差，以更快地改变人类的内部模型。经过初始阶段这种夸大的偏离后，与相同步数下的被动学习轨迹相比，人类轨迹更接近最优路径（人类和机器人动作的详细可视化见附录 A.4）。

我们还对李克特量表调查问题进行了方差分析。调查问题考察了参与者感知到的性能提升（例如：“互动结束时，我很容易控制机器人完成任务。”）以及对机器人的理解（例如：“互动结束时，我理解了机器人的物理属性。”）。综合所有问题，我们没有发现机器人策略的显著影响（拒绝 **H7**）。令我们惊讶的是，尽管参与者在教学条件下的量化 表现更好，他们却没有感知到 性能有所提升（$p = 0.689$），对机器人物理学的理解也没有提升（$p = 0.299$）。我们推测，这可能是因为参与者与每种机器人策略只互动了一分钟，因此差异难以察觉。未来，研究与机器人的长期互动将有助于阐明这种脱节现象。

8 结论

在这项工作中，我们朝着使机器人能够理解其对人类内部模型所产生影响的目标迈出了一步。为此，我们将人类学习建模为一个非线性动力系统，该系统会随着机器人能够影响的新观测而演化。我们提出了一种易于处理的方法，可根据自然体现人类学习过程的示范，推断近似的人类学习动力学；并提出了如何将近似动力学嵌入机器人规划，从而使机器人能够影响人类学习。我们的实验结果表明，机器人影响是可行的，并且能够帮助人类学习更好的内部模型。

局限性与未来工作。我们方法的优势与局限在于通过 Transformer 表示人类学习动态。作为一种通用函数逼近器，它对人类学习动力学的结构不作任何假设；事实上，我们很欣喜地发现，结果表明，无需先验假设，就有可能从真实数据中推断出有用的人类学习模型。然而，由于神经网络需要大量人类数据，因此不适用于低数据量场景，而且在人类属于分布外情况时可能会失效。另一个局限是，如果个体并非像式（2）中那样是带噪声的最优决策者，且存在特定偏差（例如近视偏差），那么 Transformer 就会学习到用于补偿这些偏差的参数；反过来，这可能导致机器人以非预期的方式影响人类。未来，我们期待结合数据驱动模型与认知科学模型在人类学习方面的优势。尽管我们的用户研究依赖于利用所有参与者的数据训练得到的“平均”人类学习动力学模型

但人类可能会表现出独特的学习方式。推断个性化学习动力学是一个令人振奋的未来研究方向，而人类预训练模型可以作为适应新用户的有用起点。最后，虽然 LQ近似使推断变得易于处理，但将其扩展到非LQ情形将有助于解锁更多应用场景（例如自动驾驶汽车）。

参考文献

[1] Henny Admoni 和 Brian Scassellati。2017。人机交互中的社交注视：综述。《人机交互期刊》6，1（2017），25–63。[2] Chris L Baker、Rebecca Saxe 和 Joshua B Tenenbaum。2009。将动作理解视为逆向规划。《认知》113，3（2009），329–349。[3] Dimitri P Bertsekas 等。2011。《动态规划与最优控制》第3版，第II卷。马萨诸塞州贝尔蒙特：Athena Scientific（2011）。[4] Maya Cakmak 和 Manuel Lopes。2012。序贯决策任务的算法教学与人类教学。载于人工智能会议。[5] Eduardo F Camacho 和 Carlos Bordons Alba。2013。《模型预测控制》。Springer科学与商业媒体。[6] Lawrence Chan、Dylan Hadfield-Menell、Siddhartha Srinivasa 和 Anca Dragan。2019。辅助型多臂老虎机。载于2019年第14届ACM/IEEE国际人机交互会议（HRI）。IEEE，354–363。[7] Min Chen、Stefanos Nikolaidis、Harold Soh、David Hsu 和 Siddhartha Srinivasa。2020。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）（2020）。[8] Anca D Dragan、Kenton CT Lee 和 Siddhartha S Srinivasa。2013。机器人运动的可理解性与可预测性。载于2013年第8届ACM/IEEE国际人机交互会议（HRI）。IEEE，301–308。[9] Anca D Dragan 和 Siddhartha S Srinivasa。2013。用于共享控制的策略融合形式体系。《国际机器人研究期刊》32，7（2013），790–805。[10] Sebastian East、Marco Gallieri、Jonathan Masci、Jan Koutník 和 Mark Cannon。2020。无限时域可微模型预测控制。国际学习表征会议（2020）。[11] Hugging Face。2022。Transformers。https://huggingface.co/docs/transformers/index[12] Noah D Goodman、Joshua B. Tenenbaum 和 ProbMods贡献者。2016。《认知的概率模型》。http://probmods.org/v2。访问日期：2022-12-9。[13] Thomas L Griffiths、Nick Chater、Charles Kemp、Amy Perfors 和 Joshua B Tenenbaum。2010。认知的概率模型：探索表征与归纳偏置。《认知科学趋势》14，8（2010），357–364。[14] Soheil Habibian 和 Dylan P. Losey。2022。鼓励人类与机器人团队互动：清晰且公平的子任务分配。《IEEE机器人与自动化快报》8（2022），188。[15] Haimin Hu 和 Jaime F Fisac。2022。降低人机交互中的主动不确定性：一种隐式对偶控制方法。WAFR（2022）。[16] Sandy H Huang、David Held、Pieter Abbeel 和 Anca D Dragan。2019。使机器人能够传达其目标。Autonomous Robots 43，2（2019），309–326。[17] Siddarth Jain 和 Brenna Argall。2019。用于辅助机器人技术中共享自主的概率人类意图识别。《ACM人机交互汇刊》（THRI）9，1（2019），1–23。[18] Michael Janner、Qiyang Li 和 Sergey Levine。2021。将离线强化学习视为一个大型序列建模问题。《神经信息处理系统进展》34（2021），1273–1286。[19] Julian Jara-Ettinger。2019。作为逆强化学习的心智理论。《行为科学最新观点》29（2019），105–110。[20] Rudolf Emil Kalman 等。1960。对最优控制理论的贡献。Bol. soc. mat. mexicana 5，2（1960），102–119。[21] Wisama Khalil 和 Etienne Dombre。2002。机器人的建模、辨识与控制。CRC出版社。[22] Kris M Kitani、Brian D Ziebart、James Andrew Bagnell 和 Martial Hebert。2012。活动预测。载于欧洲计算机视觉会议。Springer，201–214。[23] Forrest Laine、David Fridovich-Keil、Chih-Yuan Chiu 和 Claire Tomlin。2021。博弈论运动规划中的多假设交互。载于2021年IEEE机器人与自动化国际会议（ICRA）。IEEE，8016–8023。[24] Sergey Levine 和 Vladlen Koltun。2012。利用局部最优示例进行连续逆向最优控制。arXiv预印本 arXiv:1206.4617（2012）。[25] Kejun Li、Maegan Tucker、Erdem Bıyık、Ellen Novoseller、Joel W Burdick、Yanan Sui、Dorsa Sadigh、Yisong Yue 和 Aaron D Ames。2021。Roial：用于表征外骨骼步态偏好景观的感兴趣区域主动学习。载于2021年IEEE机器人与自动化国际会议（ICRA）。IEEE，3212–3218。[26] Dylan P Losey、Hong Jun Jeon、Mengxi Li、Krishnan Srinivasan、Ajay Mandlekar、Animesh Garg、Jeannette Bohg 和 Dorsa Sadigh。2022。学习潜在动作用于控制辅助机器人。《自主机器人》46, 1 (2022), 115–147。[27] David Marr. 1982.《视觉：对人类视觉信息表征与处理的计算研究》。W.H. Freeman。[28] Max Mulder. 1999.《隧道式天空显示的控制论》。(1999)。[29] Benjamin A Newman, Reuben M Aronson, Siddhartha S Srinivasa, Kris Kitani, and Henny Admoni. 2022. HARMONIC：一个用于辅助性人机协作的多模态数据集。《国际机器人研究期刊》(2022)。[30] Benjamin A Newman, Abhijat Biswas, Sarthak Ahuja, Siddharth Girdhar, Kris K Kitani, and Henny Admoni. 2020. 研究预期式机器人辅助对人类决策的影响。载于《国际社会机器人会议》。Springer, 590–603。[31] Andrew Y Ng, Stuart J Russell, 等. 2000. 逆强化学习算法。载于 ICML，第 1 卷。2。[32] Stefanos Nikolaidis, Swaprava Nath, Ariel D Procaccia, and Siddhartha Srinivasa. 2017. 人机协作中人类适应的博弈论建模。载于 2017 年 ACM/IEEE国际人机交互会议论文集。323–331。[33] Sagar Parekh, Soheil Habibian, and Dylan P Losey. 2022. RILI：稳健地影响潜在意图。arXiv 预印本 arXiv:2203.12705 (2022)。[34] Lasse Peters, David Fridovich-Keil, Vicenç Rubies-Royo, Claire J Tomlin, and Cyrill Stachniss. 2021. 从受噪声干扰的部分状态观测中推断连续动态博弈的目标。arXiv 预印本 arXiv:2106.03611 (2021)。[35] Mark Pfeiffer, Ulrich Schwesinger, Hannes Sommer, Enric Galceran, and Roland Siegwart. 2016. 预测行动以实现可预测的行动：基于最大熵模型的协作式部分运动规划。载于 2016 年 IEEE/RSJ智能机器人与系统国际会议（IROS）。IEEE, 2096–2101。[36] David Premack and Guy Woodruff. 1978. 黑猩猩有心智理论吗?《行为与脑科学》1, 4 (1978), 515–526。[37] Irene Rae, Leila Takayama, and Bilge Mutlu. 2013. 高度对机器人媒介沟通的影响。载于 ACM/IEEE国际人机交互会议。1–8。[38] Sid Reddy, Anca Dragan, and Sergey Levine. 2018. 你觉得自己要去哪儿？：从行为中推断关于动力学的信念。《神经信息处理系统进展》第 31 卷 (2018)。[39] Siddharth Reddy, Sergey Levine, and Anca D Dragan. 2020. 辅助感知：优化观测以传达状态。arXiv 预印本 arXiv:2008.02840 (2020)。[40] Dorsa Sadigh, Shankar Sastry, Sanjit A Seshia, and Anca D Dragan. 2016. 为能利用人类行为影响的自动驾驶汽车进行规划。载于《机器人学：科学与系统》。[41] Shane Saunderson and Goldie Nejat. 2019. 机器人如何影响人类：社会性人机交互中非语言沟通综述。《国际社会机器人期刊》11, 4 (2019), 575–608。[42] Sydney Y Schaefer, I ris L Shelly, and Kurt A Thoroughman. 2012. 并非重点：在任务无关条件下，不通过基于反馈的误差校正实现运动适应。《神经生理学杂志》107, 4 (2012), 1247–1256。[43] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. 近端策略优化算法。arXiv 预印本 arXiv:1707.06347 (2017)。[44] Wilko Schwarting, Alyssa Pierson, Javier Alonso-Mora, Sertac Karaman, and Daniela Rus. 2019. 自动驾驶车辆的社会行为。《美国国家科学院院刊》116, 50 (2019), 24972–24978。[45] Lei Shi, Naomi H Feldman, and Thomas L Griffiths. 2008. 使用范例模型进行贝叶斯推断。载于认知科学学会年会论文集，第 30 卷。[46] Megha Srivastava, Erdem Biyik, Suvir Mirchandani, Noah Goodman, and Dorsa Sadigh. 2022. 面向人类运动控制任务的辅助式教学。arXiv 预印本 arXiv:2211.14003 (2022)。[47] Richard S Sutton and Andrew G Barto. 2018.《强化学习：导论》。MIT出版社。[48] Ran Tian, Liting Sun, Andrea Bajcsy, Masayoshi Tomizuka, and Anca D Dragan. 2022. 通过置信度感知的博弈论人类模型为人机交互提供安全保障。载于 2022 年机器人与自动化国际会议（ICRA）。IEEE, 11229–11235。[49] Luke Tierney and Joseph B Kadane. 1986. 后验矩与边缘密度的精确近似。《美国统计学会会刊》81, 393 (1986), 82–86。[50] Tomer D. Ullman and Joshua B. Tenenbaum. 2020. 概念发展的贝叶斯模型：将学习视为构建世界模型。《发展心理学年度评论》2, 1 (2020), 533–558。https://doi.org/10. 1146/annurev-devpsych-121318-084833 arXiv:https://doi.org/10.1146/annurev-devpsych-121318-084833[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. 注意力就是你所需的一切。《神经信息处理系统进展》第 30 卷 (2017)。[52] Kevin Waugh, Brian D Ziebart, and J Andrew Bagnell. 2010. 矩阵博弈的逆相关均衡。《神经信息处理系统进展》(2010)。

[53] Yair Weiss, Eero P Simoncelli, and Edward H Adelson. 2002. 作为最优知觉的运动错觉。《自然-神经科学》5, 6 (2002), 598–604。[54] Annie Xie, Dylan P Losey, Ryan Tolsma, Chelsea Finn, and Dorsa Sadigh. 2020. 学习潜在表征以影响多智能体交互。机器人学习会议(2020)。[55] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. 2008. 最大熵逆强化学习。载于 *AAAI*，第 8 卷。美国伊利诺伊州芝加哥，1433–1438。

附录 A

A.1 推导：LQ近似下的高斯积分

这里我们推导式 (2) 中分母在 LQ近似下的闭式解。首先，回顾高斯积分：

定理：高斯积分 *[49]*。设 $M \in \mathbb{R}^{n\times n}$ 为一个对称正定矩阵，且 $x \in \mathbb{R}^n$。则：

$$\int \exp\left(-\frac{1}{2}x^\top Mx + b^\top x\right)d^n x = \sqrt{\frac{(2\pi)^n}{\det(M)}}\exp\left(\frac{1}{2}b^\top M^{-1}b\right). \quad (15)$$

定理：无限时域线性二次调节器 *[3]*。设离散时间动力学为线性，$x^{t+1} = Ax^t + Bu^t$，且代价为二次型，$x^\top Qx + u^\top Ru$。则无限时域最优剩余代价 J 和最优控制 $u^*(x)$ 为：

$$J(x) = x^\top Px \quad (16)$$

$$u^*(x) = -Kx \quad (17)$$

其中 P 是无限时域离散时间黎卡提方程（离散代数黎卡提方程）的唯一正定不动点：

$$P = A^\top PA - A^\top PB(R + B^\top PB)^{-1}B^\top PA + Q \quad (18)$$

以及反馈矩阵 $K = (R + B^\top PB)^{-1}B^\top PA$。

推导。设物理动力学为线性，奖励关于状态和控制为二次函数。假设我们已将状态-动作Q_H 函数近似为：

$$Q_\mathrm{H}(x,u) = -x^\top Qx - u^\top Ru - (x')^\top P(x') \quad (19)$$

其中 $x\prime = Ax + Bu$ 是下一个状态，P是方程（18）的解。将（19）代入人类策略的分母，可得：

$$\int \exp\left(Q\mathrm{H}(x,u)\right)du \quad (20)$$

$$= \int \exp\left(-x^\top Qx - u^\top Ru - (Ax+Bu)^\top P(Ax+Bu)\right)du \quad (21)$$

$$= \exp\left(-x^\top Qx - x^\top A^\top PAx\right)\cdot \quad (22)$$

$$\int \exp\left(-\frac{1}{2}u^\top(2R + 2B^\top PB)u + (-2x^\top A^\top PB)u\right)du \quad (23)$$

我们发现，若令 $M := 2R + 2B^\top PB$ 和 $b := -2x^\top A^\top PB$，则可以直接计算高斯积分，得到：

$$= \exp\left(-x^\top Qx - x^\top A^\top PAx\right)\sqrt{\frac{(2\pi)^m}{\det(2R + 2B^\top PB)}}\cdot \quad (24)$$

$$\exp\left((-2x^\top A^\top PB)^\top[2R + 2B^\top PB]^{-1}(-2x^\top A^\top PB)\right) \quad (25)$$

$$= \exp\left(x^\top\left[A^\top PB(R + B^\top PB)^{-1}B^\top PA - Q - A^\top PA\right]x\right)\cdot \quad (26)$$

$$\sqrt{\frac{(2\pi)^m}{\det(2R + 2B^\top PB)}}\cdot \quad (27)$$

有趣的是，我们发现指数项在括号内包含了（取负后的）式（18）中的离散代数黎卡提方程。代回 P 后，得到：

$$\int \exp\left(Q_\mathrm{H}(x,u)\right)du = \exp\left(-x^\top Px\right)\sqrt{\frac{(2\pi)^m}{\det(2R + 2B^\top PB)}}. \quad (28)$$

我们还可以通过以下方式进一步简化这个方程

$$x^\top Px = \min_u\left[x^\top Qx + u^\top Ru + (Ax+Bu)^\top P(Ax+Bu)\right] \quad (29)$$

$$= x^\top Qx + (u^*)^\top R(u^*) + (Ax+Bu^*)^\top P(Ax+Bu^*) \quad (30)$$

$$= Q_\mathrm{H}(x,u^*) \quad (31)$$

其中，u^* 是状态 x 下的最优控制。因此，我们可以得到最终的简化形式：

$$\int \exp\left(Q_\mathrm{H}(x,u)\right)du = \exp\left(-Q_\mathrm{H}(x,u^*)\right)\sqrt{\frac{(2\pi)^m}{\det(2R + 2B^\top PB)}}.$$

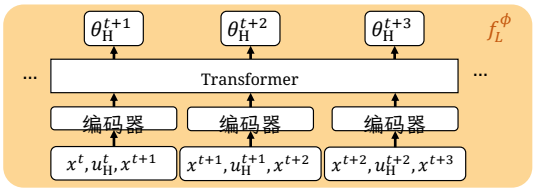


图 6：表示 f_L^ϕ 的 Transformer 架构。

A.2 关于推断人类学习动力学的基于梯度的解的详细说明

为了优化基于 Transformer 的人类学习动力学模型，我们需要计算推断目标（公式 (5) 中给出，以下记为 L）相对于神经网络参数 ϕ 的梯度：

$$\frac{\partial\mathcal{L}}{\partial\phi} = \frac{\partial\mathcal{L}}{\partial\theta_\mathrm{H}}\frac{\partial\theta_\mathrm{H}}{\partial\phi}.$$

第二个分量 $\frac{\partial\theta_\mathrm{H}}{\partial\phi}$ 是 Transformer 内部模型预测相对于神经网络权重的梯度，由于 Transformer 是可微的，因此可以直接获得。然而，第一个分量

$$\frac{\partial\mathcal{L}}{\partial\theta_\mathrm{H}} = \left[\frac{\partial\log\mathbb{P}(u_\mathrm{H}^t \mid x^t;\theta_\mathrm{H}^t)}{\partial{\theta_\mathrm{H}}^t}\right],$$

即人类策略相对于人类内部模型参数的策略梯度，是一个关键挑战。这是因为人类策略 $\mathbb{P}(u\mathrm{H} \mid x,\theta\mathrm{H})$ 通过 $Q\mathrm{H}$-值依赖于 P_{θ_H}。回想一下，P_{θ_H} 是方程（10）中离散代数黎卡提方程的解，它依赖于矩阵 A、B、Q、R 和 P_{θ_H}。无论人类内部模型参数是物理动力学 $\theta\mathrm{H} := (A,B)$，还是奖励权重 $\theta\mathrm{H} := (Q,R)$，计算人类策略梯度都需要对 DARE 函数求导，这并非显而易见。借助近期研究 [10]（该研究将 DARE 视为 (A,B,Q,R) 的隐函数），我们得到了闭式雅可比矩阵 $\frac{\partial P}{\partial A}$、$\frac{\partial P}{\partial B}$、$\frac{\partial P}{\partial Q}$ 和 $\frac{\partial P}{\partial R}$。这些矩阵的具体形式可以 __

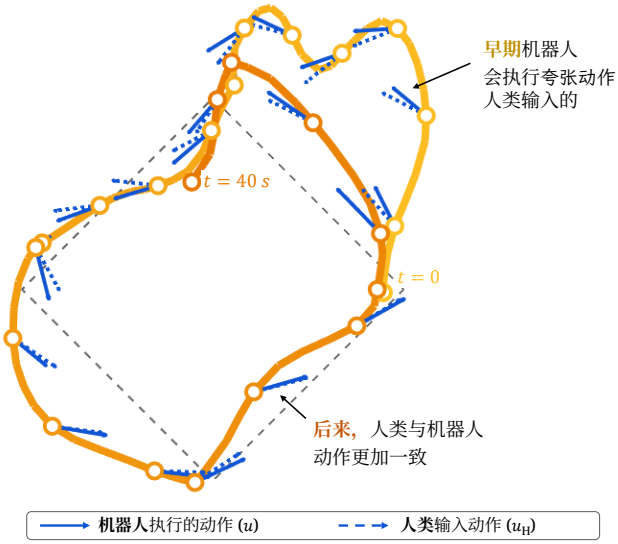


图 7：第 7 节中的一个参与者轨迹示例。蓝色矢量显示机器人执行动作（实线）和人类输入动作（虚线），采样时间步间隔为 0.5 秒。

在 [10] 的命题2中找到。因此，我们可以高效地计算$\partial\mathcal{L}/\partial\phi$，并通过基于梯度的优化推断人类的学习动力学。

A.3 训练人类学习动力学模型

为提高推断 f_L^ϕ 的可复现性，我们在此介绍架构和优化细节。在所有设置中，用于编码 $(x^t,\ u_H^t\ x^{t+1})$ 的编码器都是一个包含 3 个全连接层的多层感知机。我们采用 Hugging Face 提供的实现 [11]，使用 Transformer 编码器 [51] 表征人类学习动力学，并使用 Adam优化器训练神经网络。

在仿真实验和用户研究中，我们使用相同的 Transformer 架构来表征人类学习动态，仅针对各项任务调整输出层的大小，以便恰当地建模人类内部模型 θ_H。根据第 6.1 节，在 **月球着陆器** 任务中，输出维度为 2，表示人类正在估计的 B向量；在 **机械臂遥操作** 任务中，输出维度为 4，以表示 B 和 w的对角元素。根据第 6.2 节，在 **目标影响** 任务中，输出维度为 2，表示托盘第一个和第二个目标的概率（即人类信念）；第三个目标的概率则隐式定义为 1 减去前两个目标概率之和。而在 **偏好影响** 任务中，输出维度为 3，用于表示 $Q \in \mathbb{R}^{3\times3}$的对角项。最后，在第 7 节的用户研究中，输出维度为 4，以表示 B 和 w的对角元素。请注意，在预测 θ_H^1 时，人类的初始内部模型（θ_H^0）会在输入开头处被隐式估计。

A.4 用户研究：人类与机器人动作对齐

我们查看了用户研究数据，并研究了在我们的教学方法下，人类输入动作与机器人执行动作之间的差异。回想一下，机器人根据（13）执行动作：$u = \alpha \cdot u_R + (1-\alpha) \cdot u_H$ 其中，在用户研究期间，$\alpha = 0.5$，而u_R 则是根据我们的影响感知规划方法生成的。在图 7 中，我们绘制了一条示例参与者轨迹，以及每隔 1.5 秒的机器人执行动作（蓝色实线向量）和人类输入动作（蓝色虚线向量）。从定性角度看，研究初期，人类与机器人的动作并不一致。直观而言，由于机器人的规划目标是让人类的物理学心智模型尽快与机器人的物理模型一致，因此机器人计划夸张地执行人类输入的动作，以期尽快改变其想法。之后，我们看到，随着人类逐渐学会成为更好的遥操作员，人类与机器人的动作也变得更加一致。