即将在 2023 年 IEEE 国际可信自主系统会议（ICAA）上发表。

通过部分可观测马尔可夫决策过程树搜索进行奖励模型协调的解释

Benjamin D. Kraske*、Anshu Saksena†、Anna L. Buczak†、Zachary N. Sunberg*
*美国科罗拉多州博尔德市科罗拉多大学博尔德分校航空航天工程科学系
†美国马里兰州劳雷尔市约翰斯·霍普金斯大学应用物理实验室
*Benjamin.Kraske@colorado.edu, †Anshu.Saksena@jhuapl.edu, †Anna.Buczak@jhuapl.edu, *Zachary.Sunberg@colorado.edu

arXiv:2305.00931v1 [cs.AI] 1 May 2023

摘要——**随着人工智能（AI）算法日益广泛地应用于关键任务领域，提升用户对这些系统的信任对其成功至关重要。确保用户理解算法进行推理所依据的模型，有助于增进用户信任。本研究旨在协调算法进行在线部分可观测马尔可夫决策过程（POMDP）规划时所使用的奖励模型，与人类用户所假定的隐式奖励模型之间的差异。我们利用动作差异——即算法与用户所作决策之间的差异——来估计用户的目标，并以奖励函数的权重形式表示这些目标。**

索引词——**可解释人工智能（XAI）、部分可观测马尔可夫决策过程（POMDP）、部分可观测马尔可夫决策过程规划**

I. 引言

以序贯决策算法为形式的人工智能正日益用于解决现实世界的问题。这些算法有诸多优势，包括在许多情况下比人类更有效地推理并考虑未来结果。然而，只有当人们信任这些系统所建议的决策时，它们才能在现实应用中发挥作用。如果算法与用户的目标不同，双方作出的决策和产生的结果很可能也会不同，从而造成困惑、降低对系统的信任，甚至导致用户弃用系统。

考虑资源分配问题：有限数量的资源必须高效地分配给多项需求。此类问题的例子包括太空域感知传感器任务分配问题（有限数量的传感器需要监测多个目标），以及维修调度问题（有限数量的维修人员需要前往多个位置进行维修）。在这些情况下，我们假设用户会批准算法作出的分配决策。此时，至关重要的是，用户能够理解算法为何作出该决策，并在批准决策之前消除对模型或规划过程的任何误解，最好能实时完成。

本研究旨在协调用户与算法的部分可观测马尔可夫决策过程（POMDP）模型之间的差异，特别关注奖励模型中编码的目标。我们利用算法与用户之间的动作差异来估计用户的奖励模型，并针对用户对模型的误解提供量身定制的解释。

本材料基于约翰斯·霍普金斯大学应用物理实验室和美国海军研究办公室资助的工作。

第二节介绍可解释人工智能（XAI）和模型协调的背景。第三节概述解释问题。第四节介绍我们提出的解决方案。第五节提供一个说明性示例，第六节则进行讨论。

II. 背景与相关工作

A. 部分可观测马尔可夫决策过程

本文假设读者熟悉部分可观测马尔可夫决策过程 [1], ，其定义为由状态空间 $\mathcal{S}$、动作空间 $\mathcal{A}$、转移模型 $\mathcal{T}$、观测空间 $\mathcal{O}$、观测模型 $\mathcal{Z}$、奖励模型 $\mathcal{R}$ 和折扣因子 γ 组成的元组。

B. 可解释人工智能

可解释人工智能（XAI）旨在通过促进用户对算法模型及其结果的理解，提高算法的透明度。尽管可解释人工智能的研究大多聚焦于机器学习和黑箱方法，但对可解释规划 [2] 的关注正在不断增加。

与本文最相关的可解释人工智能领域由 Chakraborti 等人 [3]，提出。他们描述了模型协调问题：解释旨在使用户推理所依据的模型与规划算法所依据的模型保持一致。Sreedharan 等人 [4] 将模型协调（针对转移函数、奖励函数和折扣因子）应用于马尔可夫决策过程，并采用基于学习的方法来确定应解释哪些模型参数。Tabrez 等人 [5] 使用增强型部分可观测马尔可夫决策过程模型，为马尔可夫决策过程的奖励模型提供解释。Wang 等人 [6] 对部分可观测马尔可夫决策过程规划进行了解释，但对奖励的讨论主要涉及呈现结果和数值。Yadav 等人 [7] 将影响力最大化问题表述为部分可观测马尔可夫决策过程，并开展用户研究，以了解用户如何推理定义该问题的网络，为解释其解决方案奠定基础。

Israelsen 和 Ahmed [8] 回顾了人类—自主系统信任中的算法保障。具体而言，文中讨论了价值对齐或人工智能对齐，即用户目标与算法目标相一致。Bobu 等人 [9] 回顾了使人类与机器人的任务表征相一致的方法，包括奖励学习。Yuan 等人 [10] 通过协作博弈建模，提出了一种双向的人机价值对齐方法。我们的工作旨在了解用户目标，以便

© 2023 IEEE。允许个人使用本材料。对于任何其他用途，无论是在当前还是未来的媒体中，包括为广告或宣传目的转载/再版本材料、创作新的合集作品、转售或重新分发至服务器或列表，或在其他作品中重用本作品的任何受版权保护部分，均须事先获得 IEEE 的许可。

能够针对当前算法目标提供相关解释。

C. 部分可观测马尔可夫决策过程逆强化学习

逆向强化学习 (逆向强化学习) 旨在从专家轨迹中学习目标，这些目标在马尔可夫决策过程的语境下通过奖励函数表达。大量研究已经探讨了部分可观测马尔可夫决策过程逆向强化学习 [11]–[14]。Atrash and Pineau [15]提出了一种强化学习框架，通过向最优预言机查询动作，为真实奖励分布的估计提供信息。我们的工作与这些研究的不同之处在于，我们基于单个动作和信念状态对奖励函数进行粗略估计。尽管原则上这些逆向强化学习方法可以应用于我们的问题，但我们尚未将其作为对比方法。

III. 问题陈述

确定向用户呈现的最少解释数量，使其模型得到更新，同时尽量减少信息交换，是模型协调中的一项关键挑战 [4]。我们的工作旨在解决目标方面的模型差异，这些目标通过 POMDP 奖励函数来表达。

在许多情况下，部分可观测马尔可夫决策过程的奖励函数 R 可以自然地表示为特征向量 $\boldsymbol{\beta}(s,a)$ 与权重 ϕ，使得 $R(s,a) = \phi^T \boldsymbol{\beta}(s,a)$。我们假设用户理解奖励特征 $\boldsymbol{\beta}(s,a)$，但对每个奖励特征的重视程度（或权重）不同，ϕ。我们的工作旨在估计用户权重，然后利用这些权重向用户提供简洁的解释，从而有效更新用户模型。更正式地说，给定部分可观测马尔可夫决策过程模型 m、信念状态 b_τ、规划算法的动作$a_{\phi_a,\tau}$以及用户提出的替代动作 $a_{\phi_h,\tau}$，我们的目标是ˆ 找到用户权重 ϕ_h的估计值 $\hat{\phi}_h$ 。

IV. 解决方案

我们将给定时间步上的用户提议动作作为估计用户奖励权重 ϕ_h的基础，并采用一种简单的类逆向强化学习方案。利用该动作，我们通过寻找 ϕ 来估计用户奖励权重，使用户提议的动作具有高于算法动作的估计值（图 1）。

A. 约束优化

我们基于单个时间步的反馈来生成解释，并将其称为单次方法。这种方法只需一个用户提出的动作，就能解释其中的差异，而不必使用多个时间步或多条轨迹中的用户提出的动作。优化问题的目标是找到一个奖励加权值，ˆϕ_h，使人类动作 a_h 至少与算法选择的动作 a_a 同样好，同时尽可能接近原始奖励加权值 ϕ_a。该优化问题的数学形式如下：

$$\begin{aligned} &\underset{\hat{\phi}_h}{\text{minimize}} \quad \|\hat{\phi}_h - \phi_a\|_1 \\ &\text{subject to } Q^{\hat{\phi}_h}(b_\tau, a_{\phi_h,\tau}) \geq Q^{\hat{\phi}_h}(b_\tau, a^*_{\phi_a,\tau}) \\ &\qquad\qquad \hat{\phi}_h \geq 0 \end{aligned} \tag{1}$$

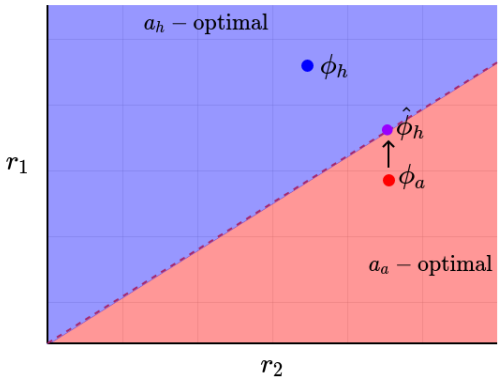


图 1. 使用动作差异估计 ϕ_h

其中，b_τ 是给定时间步 τ 的信念状态，$Q^{\phi_h}(b,a)$ 是信念状态-动作价值（根据人类奖励加权评估），$a_{\phi_h,\tau}$ 是时间步 τ 的用户提出的动作，$a^*_{\phi_a,\tau}$ 是时间步 τ 下算法奖励加权对应的最优动作。

B. 优化

通过对约束违反进行惩罚，将约束优化问题 (1) 转化为无约束优化问题，从而得到目标函数 U：

$$\begin{aligned} &U = \|\hat{\phi}_h - \phi_a\|_1 + wL_Q \\ &\text{where } L_Q = -\max(|Q^{\hat{\phi}_h}(b_\tau, a^*_{\phi_a,\tau}) - Q^{\hat{\phi}_h}(b_\tau, a_{\phi_h,\tau})|, 0) \end{aligned} \tag{2}$$

其中，w 是松弛中的权重变量，理想情况下应将其设置为使原始约束不被违反。

该松弛优化问题使用交叉熵方法 [16], 求解，并限制输出，使得 ϕ 的每个元素都大于 0。交叉熵提供了一种直接且无需梯度的方法，用于优化上述损失函数；这些损失函数依赖于 DESPOT 规划器对价值的估计，而这些估计是实时计算的。

C. 部分可观测马尔可夫决策过程解决方案

为了评估式 (2) 中定义的损失函数，需要估计 ϕ 下各个动作的 Q 值。我们使用在线部分可观测马尔可夫决策过程求解器 DESPOT [17], 来获得 Q 值估计。这样做的优势在于，我们的方法能够扩展到状态空间较大的部分可观测马尔可夫决策过程，而这类问题只能通过在线方式求解。

V. 示例说明

A. 部分可观测马尔可夫决策过程的形式化

我们将以下问题表述为一个资源分配问题，它为测试决策解释提供了一个有效示例，尤其适用于那些对潜在用户而言有悖直觉的决策。暖通空调（供暖、通风和空调）维修调度问题涉及派遣维修人员前往多个位置之一，这些位置可能处于“正常”状态，也可能发生暖通空调故障（机械故障、电气故障或冷却剂故障）。位置并非始终可用，因为客户

可能并不总是在家。派遣维修人员需要付出成本，如果故障未能在规定时间内修复，则会受到惩罚。该问题表述如下： $V = 5$ 表示可用性信息可获取的时段， $N = 3$ 表示位置数量，$R = 2$ 表示维修人员数量，$\mathrm{T} = 16$表示问题的时间范围。

状态空间

状态空间 $\mathcal{S}$ (3) 是位置状态（包括位置状态上次变更的时间）、位置可用性和时间步的笛卡尔积。

$$\begin{aligned}&\mathcal{S} = \text{Status} \times \text{Availability} \times [1, \mathrm{T}]\\&\text{Status} = \{\{\mathsf{s}_i, \mathsf{t}_i\}_{i=1}^{N} \mid \mathsf{s}_i \in \{\text{Ok}, \text{Mech.}, \text{Elec.}, \text{Cool.}\},\\&\mathsf{t}_i \in [1, \mathrm{T}]\}\\&\text{Availability} = \{\{\mathsf{al}_{vi}\}_{i=1}^{N} \mid \mathsf{al}_{vi} \in \{\text{Avail}, \text{NA}\},\\&v \in [1, V]\}\end{aligned} \tag{3}$$

动作空间

动作空间 $\mathcal{A}$ (4) 是一个长度为 R 的元组，其中每个条目对应维修人员被派往的位置（位置 0 表示不派出维修人员）。

$$\mathcal{A} = \{\{\mathsf{a}_r\}_{r=1}^{R} \mid \mathsf{a}_r \in [0, n]\} \tag{4}$$

状态转移分布

状态转移模型由三个部分组成，分别对应状态、可用性以及问题的整体时间步。处于某种故障状态（机械、冷却液、电气。）的位置，除非在其可用时将维修人员派往该位置，否则会保持该状态。维修人员 1 专门负责冷却液维修。如果将其派往某个位置，则机械故障以 0。8 的概率得到解决，电气故障以 0。9 的概率得到解决，冷却液故障则以 1。0 的概率得到解决。如果将维修人员 2 派往某个位置，则任何故障都以 0。9 的概率得到解决。位置状态以 0。7 的概率保持为正常，并以 0。1 的概率转变为每一种故障状态。如果位置状态发生变化，状态中的时间步部分将重置为当前时间步。

观测空间

观测空间 $\mathcal{O}$ 定义见式（5）。

$$\begin{aligned}&\mathcal{O} = \text{Status} \times \text{Availability} \times [1, \mathrm{T}]\\&\text{Status} = \{\{\mathsf{s}_i\}_{i=1}^{N} |_i \in \{\text{Ok}, \text{Mech.}, \text{Elec.}, \text{Cool.}\}\}\\&\text{Availability} = \{\{\mathsf{al}_{vi}\}_{i=1}^{N} \mid \mathsf{al}_{vi} \in \{\text{Avail}, \text{NA}\},\\&v \in [1, V]\}\end{aligned} \tag{5}$$

观测分布

观测模型中的状态部分会提供有关状态的含噪观测。如果某个位置正常，则以 0.7 的概率提供正常观测，并以各 0.1 的概率提供错误的故障观测。如果某个位置处于故障状态，则以 0.5 的概率准确观测到该故障，以 0.1 的概率错误地观测为正常，并以各 probability0.2 的概率错误地观测为另外两种故障。故障开始的时间步无法直接观测。位置的可用性以及问题时间步都是完全可观测的。

奖励

奖励函数使用加权项 ϕ定义：

$$\begin{aligned}&R(s, a) = \phi^T \boldsymbol{\beta}(s, a)\\&\text{where } \boldsymbol{\beta}(s, a) =\\&[\bar{R}_{L1}(s, a), ..., \bar{R}_{Ln}(s, a), \bar{R}_{W1}(s, a), ..., \bar{R}_{Wr}(s, a)]\\&\bar{R}_{Ln}(s, a) = \begin{cases} r_{l_n} \text{if Loc. } n \text{ status is fault for } x_{l_n} \\ \quad \text{timesteps} \\ 0 \text{ otherwise} \end{cases}\\&\bar{R}_{Wr}(s, a) = \begin{cases} r_{w_r} \text{ if send repairperson } r \\ 0 \text{ otherwise} \end{cases}\end{aligned} \tag{6}$$

其中， x_{l_n} 和 r_{l_n} （负数）分别为每个 N 位置设定，r_{w_r} （负数）则为每位 R 维修人员设定。默认情况下，ϕ 项均为 1，但可以调整这些项，以改变位置惩罚和维修人员工资的相对权重。所讨论示例问题的奖励参数为：$r_{l_1} = -250$、$r_{l_{2,\,3}} = -125$、$x_{l_{1,\,2,\,3}} = 3$、$r_{w_1} = -5$和$r_{w_2} = -4$。

B. 优化示例

作为一次性奖励协调的演示，我们考察仿真中的一个时间步（如图 2 所示），其中使用次优近似算法来获得解。算法奖励加权 ϕ_a 定义为 $\phi_a = [1, 1, 1, 1, 1]$。

在此 ϕ_a 下，时间步 5 执行动作 (1，1)，将两名维修人员派往位置 1。用户可能会想知道，为什么算法没有将维修人员更均匀地分配到各个位置（这些位置均发生了故障），并提出动作 (2，1)，将维修人员 1 派往位置 2，将维修人员 2 派往位置 1。给定用户动作、算法动作以及该时间步的信念状态后，执行一次性优化程序，估计加权空间边界上的一个点 ϕ ，在该点处 $Q^{\phi}(b, a_a) = Q^{\phi}(b, a_h)$（见图 1）。随后返回估计得到的人类奖励加权 $\hat{\phi}_h$ ，可将其与 ϕ_a 进行比较。

$$\hat{\phi}_h = [0.680, 1.007, 1.000, 0.999, 1.000]$$

其中，前三个元素分别对应三个位置的惩罚权重，最后两个元素对应维修人员工资的权重。

比较这两组权重后，位置 1 的权重值降至 ϕ_a 中相应数值的约 70%。权重上的这种差异，现可作为向用户简要解释部分问题目标的依据。向用户呈现这些信息有多种可能的方式。其中一种解释方法是提醒用户，按照目前的设定，位置 1 的惩罚高于其他位置，同时传达估算出的权重：

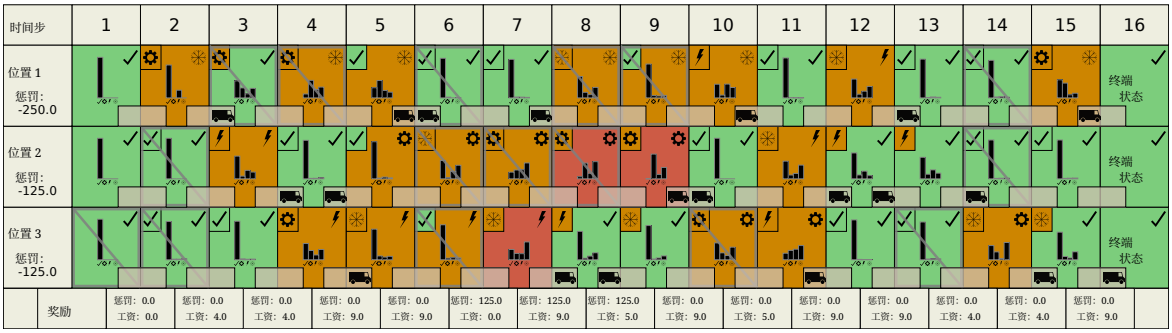


图 2. 仿真可视化。请注意，在时间步 7、8 和 9 产生的惩罚（以红色显示）。观测显示在每个单元格的左上方，真实状态显示在右上方。状态信念显示在中央，执行的动作显示在右下方。

“您似乎认为位置 1 的惩罚权重只有算法所设权重的 70%。"

根据这些信息（用户可能知道，也可能不记得），我们可以纳入用户对这些信息是否有用的反馈，并根据需要重复这一过程，用户也可以再次提出其他动作。虽然这不在本文的研究范围内，但如果需要用户就问题目标提供反馈，也可以利用这一过程更新算法的权重。

VI. 结论

上述示例为我们提出的部分可观测马尔可夫决策过程模型协调方法提供了概念验证。然而，还需要进行更全面的评估。这项工作以及一般可解释人工智能方法面临的一项挑战，是开发适当的评估方法。我们后续的工作将聚焦于评估和改进本文所述工作的方法。

此外，与许多逆向强化学习问题一样，奖励权重估计问题也存在欠规范性。对于给定的信念状态，可能有许多不同的奖励权重会导致相同的动作。我们提出的优化问题通过最小化 ϕ_a 与 ϕ_h之间的距离来避免这一问题。然而，仅凭一个动作和一个信念状态，很难准确估计用户的真实奖励权重 ϕ_h。 ϕ_h 并非在所有情况下都能很好地近似 ϕ_h，尤其是在 ϕ_h 与 ϕ_a 相差较大时（见图 1）。提高 ϕ_h的准确性（可能通过多轮解释和反馈来实现）是未来的研究方向。

致谢

作者感谢约翰斯·霍普金斯大学应用物理实验室的 Caroline Rogers 和 Michael Burke 提供意见与支持。

参考文献

[1] L. P. Kaelbling、M. L. Littman 和 A. R. Cassandra，“部分可观测随机领域中的规划与行动”，人工智能，第 101 卷，第 1–2 期，第 99–134 页，1998 年。[2] T. Chakraborti、S. Sreedharan 和 S. Kambhampati，“可解释自动规划与决策领域的新兴趋势”，载于第 *297* 2020，*20*，4803 4811，*th Int. Joint Conf. Artif. Intell., IJCAI*-C. Bessiere 编，第 – 页，综述专题。

[3] T. Chakraborti、S. Sreedharan、Y. Zhang 和 S. Kambhampati，“将计划解释作为模型协调：超越独白式解释”，载于第 *26* 届国际人工智能联合会议论文集，2017 年 8 月，第 156–163 页。[4] S. Sreedharan、A. O. Hernandez、A. P. Mishra 和 S. Kambhampati，“无模型模型协调”，载于第 *28* 届国际人工智能联合会议论文集，*IJCAI-19*，2019 年 7 月，第 587–594 页。[5] A. Tabrez、S. Agrawal 和 B. Hayes，“基于解释的奖励指导：通过强化学习提升人类表现”，载于*2019* 年第 *14* 届 *ACM/IEEE* 人机交互国际会议（*HRI*），2019 年 3 月，第 249–257 页，ISSN：2167-2148。[6] N. Wang、D. V. Pynadath 和 S. G. Hill，“人机团队中的信任校准：比较自动生成的解释”，载于*2016* 年第 *11* 届 *ACM/IEEE* 人机交互国际会议（*HRI*），2016 年 3 月，第 109–116 页，ISSN：2167-2148。[7] A. Yadav、H. Chan、A. Jiang、E. Rice、E. Kamar、B. Grosz 和 M. Tambe，“用于协助无家可归者收容所的部分可观测马尔可夫决策过程——计算与部署挑战”，载于自主智能体与多智能体系统，《计算机科学讲义》系列，N. Osman 和 C. Sierra 编。Cham：Springer International Publishing，2016 年，第 67–87 页。[8] B. W. Israelsen 和 N. R. Ahmed，“‘Dave……我可以向你保证……一切都会好起来的……’人类与自主系统信任关系中算法保障的定义、论证与综述”，《*ACM*计算调查》，第 51 卷，第 6 期，第 1–37 页，2019 年 11 月。[9] A. Bobu、A. Peng、P. Agrawal、J. Shah 和 A. D. Dragan，“协调机器人与人类的表征”，2023 年 2 月，arXiv:2302.01928 [cs]。[10] L. Yuan、X. Gao、Z. Zheng、M. Edmonds、Y. N. Wu、F. Rossano、H. Lu、Y. Zhu 和 S.-C. Zhu，“原位双向人机价值对齐”，《科学·机器人学》，第 7 卷，第 68 期，文章编号 eabm4183，2022 年 7 月，出版方：美国科学促进会。[11] J. Choi 和 K.-E. Kim，“部分可观测环境中的逆强化学习”，《机器学习研究杂志》，第 12 卷，第 21 期，第 691–730 页，2011 年。[12] H. R. Chinaei 和 B. Chaib-Draa，“面向部分可观测领域的逆强化学习算法及其在医疗对话管理中的应用”，载于*2012* 年第 *11* 届机器学习与应用国际会议，第 1 卷，2012 年 12 月，第 144–149 页。[13] H. Chinaei 和 B. Chaib-draa，“对话部分可观测马尔可夫决策过程组件（第二部分）：学习奖励函数”，国际语音技术杂志，第 17 卷，第 4 期，第 325–340 页，2014 年 12 月。[14] F. Djeumou、M. Cubuktepe、C. Lennon 和 U. Topcu，“部分信息下的任务引导逆强化学习”，自动规划与调度国际会议论文集，第 32 卷，第 53–61 页，2022 年 6 月。[15] A. Atrash 和 J. Pineau，“用于定制人机接口的贝叶斯强化学习方法”，载于第 *14* 届智能用户界面国际会议论文集。美国佛罗里达州萨尼贝尔岛：ACM，2009 年 2 月，第 355–360 页。[16] R. Y. Rubinstein 和 D. P. Kroese，交叉熵方法：组合优化、蒙特卡洛模拟与机器学习的统一方法。Springer，2004 年，第 133 卷。[17] N. Ye、A. Somani、D. Hsu 和 W. S. Lee，“DESPOT：带正则化的在线部分可观测马尔可夫决策过程规划”，人工智能研究杂志，第 58 卷，第 231–266 页，2017 年 1 月。