Active Inference through Incentive Design in Partially Observable Markov Decision Processes

Extended Abstract

Xinyi Wei
University of Florida
Gainesville, United States
weixinyi@ufl.edu

Chongyang Shi
University of Florida
Gainesville, FL, United States
c.shi@ufl.edu

Shuo Han
University of Illinois Chicago
Chicago, IL, United States
hanshuo@uic.edu

Ahmed Hemida
DEVCOM Army Research Laboratory
Adelphi, MD, United States
ahmed.h.hemida.ctr@army.mil

Charles A. Kamhoua
DEVCOM Army Research Laboratory
Adelphi, MD, United States
charles.a.kamhoua.civ@army.mil

Jie Fu
University of Florida
Gainesville, FL, United States
fujie@ufl.edu

ABSTRACT

Active inference refers to a class of methods that influence or control observed information to minimize uncertainty about latent or unknown variables. In this paper, we study a class of active inference problems in which an agent (the leader), with only partial observations, seeks to infer the unknown type of another agent (the follower), whose interactions with a dynamic environment are modeled as a Markov decision process (MDP). Different follower types are characterized by distinct dynamics, reward functions, or both, and each follower acts optimally to maximize its own reward. To improve inference accuracy and efficiency under imperfect observations, we introduce the paradigm of Active Inference through Incentive Design, wherein the leader strategically offers side payments (incentives) to elicit diverging observable behaviors from different follower types. This formulation leads to a leader–follower game in which the leader balances the trade-off between incentive cost and information gain, quantified by the entropy of the posterior distribution over follower types. We show that the resulting bi-level optimization problem can be reduced to a single-level one by leveraging the softmax temporal consistency between followers' policies and value functions. This reduction enables an efficient first-order, gradient-based algorithm, where gradients are computed using observable operators from hidden Markov models. Experimental results in stochastic gridworld environments demonstrate that the proposed method significantly improves both the accuracy and efficiency of intent inference compared to systems without incentive mechanisms.

KEYWORDS

Active inference, leader-follower game, incentive design, partially observable Markov decision processes

ACM Reference Format:
Xinyi Wei, Chongyang Shi, Shuo Han, Ahmed Hemida, Charles A. Kamhoua, and Jie Fu. 2026. Active Inference through Incentive Design in Partially Observable Markov Decision Processes: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/

1 INTRODUCTION

The ability to recognize an agent's actions, plans, and goals is fundamental in many AI tasks [7]. While extensive research has addressed the problems of identifying what the agent is doing now, what the agent wants to achieve, and how the agent plans to achieve it, a complementary question remains largely unexplored: *How can we design environments that amplify differences in observed agent behavior?* This question is closely related to *active inference*, a framework in which agents act to influence observed information in order to minimize uncertainty about latent or unknown states. For instance, a teacher may wish to design an exam that more clearly distinguishes students of different proficiency levels. Similarly, an organization may seek to construct monitoring environments that make it easier to distinguish between normal users (e.g., employees) and malicious intruders.

We study this active inference problem through the lens of incentive design. Incentive design [1, 2], also referred to as the *principal-agent* or *leader-follower* game, studies the problem where a planner or leader aims to optimize system performance while anticipating and accounting for the active interactions of multiple users or followers. This work connects active inference and incentive design as follows. Consider a setting where a leader agent observes the behavior of a follower agent only imperfectly and seeks to infer the follower's unknown type. The follower's interaction with the dynamic environment is modeled as a Markov decision process (MDP). Different follower types are characterized by distinct dynamics, discount factors, or reward functions, and each follower acts optimally to maximize its own reward. While each follower has perfect observations within its own MDP, the leader receives only imperfect and noisy observations of the follower's activities. To improve inference accuracy within a finite time horizon, the leader strategically designs an incentive policy to offer side payments (additional rewards) within the environment, influencing the followers' best responses and amplifying the divergence in observable behaviors across different follower types. These incentives,

however, incur a cost to the leader's own payoff. The leader calibrates the incentive policy to minimize both uncertainty about the follower's type and the cost of providing incentives.

2 INCENTIVE DESIGN FOR ACTIVE INFERENCE

Definition 1. *Given a collection of policies, referred to as a* policy profile *of followers, $\boldsymbol{\pi} = [\pi_i]_{i\in\mathcal{T}}$, and the set of leader's observation functions for followers, $\{E_i, i \in \mathcal{T}\}$. The following hidden Markov model (HMM) can be constructed:*

$$\mathcal{M}(\boldsymbol{\pi}) = \langle S \times \mathcal{T}, \mathbf{P}_{\boldsymbol{\pi}}, O, \mathcal{E}, \mu_0 \rangle$$

- $S \times \mathcal{T}$ *is the augmented state space. Each state (s, i) includes a state in the follower's Markov decision process (MDP) and a type of the follower.*
- $\mathbf{P}_{\boldsymbol{\pi}} : S \times \mathcal{T} \to \mathcal{D}((S \times \mathcal{T}))$ *is defined by*

$$\mathbf{P}_{\boldsymbol{\pi}}((s', j)|(s, i)) = \begin{cases} \sum_{a\in A} P_i(s'|s, a)\pi_i(s, a) & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

  In other words, at the state (s, i), follower i will take an action by following his policy $\pi_i(s)$, and the type does not change.
- $\mu_0 \in \mathcal{D}((S\times\mathcal{T}))$ *is the initial state distribution. For all $(s, i) \in S \times \mathcal{T}$, $\mu_0(s, i) = \mu_i(s)\mathbb{P}(\mathbf{T} = i)$. where $\mathbf{T}$ is a random variable representing the estimated type of the follower. $\mathbb{P}(\mathbf{T})$ is the prior distribution over possible types.*
- O *is a finite set of observations.*
- $\mathcal{E} : S \times \mathcal{T} \to \mathcal{D}(O)$ *is the observation function, defined by $\mathcal{E}(o|(s, i)) = E_i(o|s)$, which is the probability of observing o when agent i at the state s.*

Let O_t denote a random variable representing the observation at time t, and let o_t be a specific realization of this random variable. We denote the posterior estimate of the type $\mathbf{T}$ given an observation sequence $o_{0:T}$ under a policy profile $\boldsymbol{\pi}$ as $\mathbb{P}_{\boldsymbol{\pi}}(\mathbf{T}|O_{0:T} = o_{0:T})$. Next, we define the planning objective—Shannon conditional entropy.

Definition 2. *Let $Y \coloneqq O_{0:T}$. The conditional Shannon entropy of the agent's type given the observations is defined by,*

$$\begin{aligned} H(\mathbf{T} \mid Y, \mathcal{M}(\boldsymbol{\pi})) &= \sum_{y\in\mathcal{Y}} \mathbb{P}_{\boldsymbol{\pi}}(y) H(\mathbf{T}|Y = y, \mathcal{M}(\boldsymbol{\pi})) \\ &= -\sum_{i\in\mathcal{T}} \sum_{y\in\mathcal{Y}} \mathbb{P}_{\boldsymbol{\pi}}(i, y) \log \mathbb{P}_{\boldsymbol{\pi}}(i|y), \end{aligned} \tag{1}$$

where y is a sample observation, and $\mathcal{Y}$ is a set of all finite observation sequences of length T.

Problem 1 (Active inference with incentive design). *Assuming the followers always best responds to the side payments by taking an optimal policy in the follower's MDP $M_i(x)$, the leader's incentive design for active inference problem is the following bi-level optimization problem:*

$$\begin{aligned} \underset{x\in\mathcal{X}}{\text{minimize}} \quad & H(\mathbf{T}|O_{0:T}, \mathcal{M}(\boldsymbol{\pi}^*(x))) + \lambda h(x) \\ \text{subject to} \quad & \pi_i^*(x) \in \underset{\pi\in\Pi}{\operatorname{argmax}}\, V_i(\mu_i, R_i(x), \pi), \forall i \in \mathcal{T}. \end{aligned} \tag{2}$$

where $\boldsymbol{\pi}^*(x) = [\pi_i^*(x)]_{i\in\mathcal{T}}$ is a policy profile consisting of the best responses of followers given side payments x, $h\colon \mathcal{X} \to \mathbf{R}_+$ is a side payment cost function, and $\lambda > 0$ is the regularization factor.

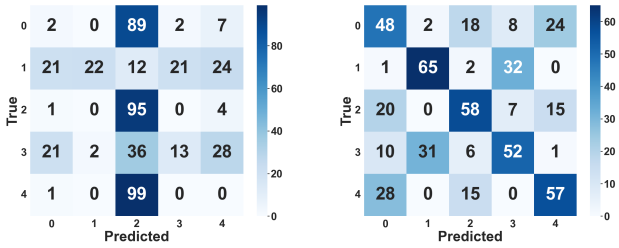
We assume that the follower employs an entropy-regularized optimal policy [5]. Based on this assumption, the best response of each follower is unique and the original bi-level optimization (2) can be reduced to a single-level optimization problem [3, 4, 6]. Then, we employ hypergradient descent algorithms to compute the leader's optimal design, given the followers best responses.

3 EXPERIMENTS

We validate the convergence of the algorithm and evaluate its efficiency in distinguishing between different agent types within a randomly generated 10×10 gridworld. The sensor can detect if an agent is within its FoV, with a false negative rate 0.05. We consider five distinct agent types with different transition dynamics and reward function. We set the regularization factor $\lambda = 0.1$. We randomly select 20 states where the leader can assign side payments across all state-action pairs for those states. The side payment cost function is defined as $h(x) = \sum_{s,a} x_{s,a}$ where the side payment $x_{s,a} \geq 0$ for each sampled state-action pair (s, a).

The initial conditional entropy $H(\mathcal{T}|Y_{0:T}; \mathcal{M}(\boldsymbol{\pi}^*(x_0)))$ with $x_0 = \vec{0}$ is 2.0457, which is close to the maximal entropy $-\log_2 \frac{1}{5} = 2.32$ derived from a uniform distribution. When the algorithm converges, the total side payment cost $h(x)$ is 5.2590 and the entropy $H(\mathcal{T}|Y_{0:T}; \mathcal{M}(\boldsymbol{\pi}^*(x^*))) = 0.8652$.

We compare the inference accuracy to a baseline which uses maximum likelihood goal recognition given the observed trajectories without side payments. We compare confusion matrices without side payments and with the computed optimal side payments, shown in Figure 1. The average true positive rate over five experiments is 28.73% (SE 0.0067, 95% CI 27.00–30.47%) without side payments and 57.12% (SE 0.0061, 95% CI 55.43–58.81%) with side payments. It is worth noting that these optimal side payments are not solely designed to minimize entropy, as the objective function is also regularized by the cost of the side payments.



(a) Confusion matrix with no side payments. **(b) Confusion matrix with the optimal incentive design.**

Figure 1: Performance of active inference: comparison of confusion matrices without and with the optimal incentive design.

ACKNOWLEDGMENTS

Research was sponsored by Air Force Office of Scientific Research under award number FA9550-21-1-0085, Army Research Office under Grant Number W911NF-22-1-0166, and NSF under award #2207759. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.

REFERENCES

[1] Patrick Bolton and Mathias Dewatripont. 2005. Contract Theory. *MIT Press Books* 1 (2005). https://ideas.repec.org//b/mtp/titles/0262025760.html Publisher: The MIT Press.

[2] Yu-Chi Ho, P Luh, and Ramal Muralidharan. 1981. Information structure, Stackelberg games, and incentive controllability. *IEEE Trans. Automat. Control* 26, 2 (1981), 454–460.

[3] Bo Liu, Mao Ye, Stephen Wright, Peter Stone, and Qiang Liu. 2022. Bome! bilevel optimization made easy: A simple first-order approach. *Advances in neural information processing systems* 35 (2022), 17248–17262.

[4] Haoxiang Ma, Shuo Han, Ahmed Hemida, Charles Kamhoua, and Jie Fu. 2024. Adaptive Incentive Design for Markov Decision Processes with Unknown Rewards. (2024). OpenReview.

[5] Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. 2017. Bridging the gap between value and policy based reinforcement learning. *Advances in neural information processing systems* 30 (2017).

[6] Vinzenz Thoma, Barna Pásztor, Andreas Krause, Giorgia Ramponi, and Yifan Hu. 2024. Contextual bilevel reinforcement learning for incentive alignment. *Advances in Neural Information Processing Systems* 37 (2024), 127369–127435.

[7] Franz A Van-Horenbeke and Angelika Peer. 2021. Activity, plan, and goal recognition: A review. *Frontiers in Robotics and AI* 8 (2021), 643010.