Continuous Multi-Attribute Recognition for Agent Behaviour Understanding

Sheryl Mantik[1] , **Michael Dann**[1] , **Huong Ha**[1] , **Minyi Li**[1] and **Julie Porteous**[1]
[1]RMIT University
sheryl.mantik@student.rmit.edu.au, michael.dann@rmit.edu.au, huong.ha@rmit.edu.au,
minyi.li2@rmit.edu.au, julie.porteous@rmit.edu.au

Abstract

Understanding underlying agent attributes such as goals, preferences, beliefs, and ability level is key to explaining decision-making and predicting behaviour in complex environments. While goal recognition (GR) has produced effective methods for inferring goals, recent work has begun to explore reinforcement learning (RL)-based approaches to attribute recognition (AR), which generalises GR by inferring a wider range of agent attributes beyond goals. Existing frameworks treat attributes as discrete variables and typically require separate models for each attribute. In reality, many agent attributes vary continuously rather than as discrete values, making continuous representations crucial for capturing subtle variations in behaviour. We introduce a multivariate RL-based AR framework that jointly infers multiple continuous-valued attributes from observed behaviour using a single attribute-conditioned policy and continuous probabilistic inference. Our experiments demonstrate that the proposed framework achieves stable and fine-grained inference, offering improved flexibility and generality compared to existing approaches.

1 Introduction

Understanding underlying agent attributes such as goals, preferences, beliefs, and ability level, is essential for explaining decision-making and for reliably predicting behaviour in complex environments [Albrecht and Stone, 2018]. Goal Recognition (GR) has produced effective methods for inferring an agent's goal from observed behaviour [Kautz and Allen, 1986; Ramírez and Geffner, 2010; Amado *et al.*, 2022], but the goal alone often provides an incomplete explanation of why agents act as they do. To address this limitation, Attribute Recognition (AR) generalises GR by inferring a broader range of latent agent attributes beyond goals [Mantik *et al.*, 2025]. Recent GR advances have incorporated Reinforcement Learning (RL) to improve adaptability in stochastic domains and reduce reliance on manual domain specification [Amado *et al.*, 2022]. Building on this approach, RL-based AR methods learn policies to infer diverse agent attributes from observed behaviour [Mantik *et al.*, 2025]. However, existing RL-based AR frameworks treat predicted attribute values as discrete variables and typically perform univariate inference, requiring separate RL policies for each attribute, creating two fundamental limitations.

First, discrete attribute representations fail to capture the inherently continuous nature of many agent attributes. Consider a videogame where a player navigates a map to collect resources and craft an item. Although their goal might be to obtain a specific rare item, their decision-making is shaped by their ability level, which varies along a continuous spectrum. As ability increases, players may progressively favour shorter and riskier routes that demand more precise timing, while lower-ability players prefer longer but safer paths to avoid enemies. To an outside observer, these behaviours could appear as if the players have different goals, even though the underlying goal remains the same. The difference lies in the player's ability level, a continuous attribute that shapes behaviour. Discretising such attributes into categories (e.g., "novice", "expert") misrepresents the problem by obscuring the nuanced ways in which incremental changes in ability affect decision-making.

Second, attributes interact and jointly shape behaviour, so inferring them independently loses critical information. Mantik *et al.* showed that ability and goal inference are tightly coupled. The same goal pursued differently at different ability levels becomes indistinguishable from a goal change when attributes are treated in isolation [Mantik *et al.*, 2025]. Worse, scaling independent inference to continuous, high-dimensional attribute spaces triggers a combinatorial explosion in required policies. A unified framework that infers all attributes jointly within a single model is therefore not merely convenient — it is necessary.

Our main contribution is a multivariate continuous RL-based AR framework that enables joint inference of **continuous attributes** within a **single policy conditioned on multiple attributes**. This addresses two key challenges: avoiding the need to train separate policies for every combination, and performing inference over continuous rather than discrete attribute values, requiring smooth density fitting over transformed attribute spaces. To the best of our knowledge, this is the first approach in both GR and AR to infer continuous attribute values. We demonstrate that our approach maintains inference accuracy while improving flexibility and generality compared to discrete baselines.

2 Attribute Recognition via RL

In this work, we focus on an extension of Goal Recognition (GR) [Kautz and Allen, 1986; Ramírez and Geffner, 2010; Amado *et al.*, 2022] called Attribute Recognition (AR) [Mantik *et al.*, 2025]. The AR problem concerns inferring specified attributes of an agent based on its observed behaviour.

Formally, the AR problem is defined as follows:

Definition 1 (Attribute Recognition)**.** *Attribute Recognition is a tuple $AR = \langle \mathfrak{D}, \mathcal{A}, O, Prob \rangle$ consisting of: the domain model $\mathfrak{D}$; the attribute space $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_n$, formed by the Cartesian product of n attribute types; a sequence of observations $O = \langle o_1, o_2, \ldots, o_m \rangle$; and a probability distribution $Prob$ over $\mathcal{A}$.*

The domain model $\mathfrak{D}$ encodes the agent's environment and is acquired by framing the task as a Markov Decision Process (MDP) $\mathcal{M} = \langle S, A, T, \gamma, R \rangle$, which will be explained more in Section 2.1. The attribute space $\mathcal{A}$ defines the range of possible values for each attribute type $\mathcal{A}_i$, which may be discrete (e.g., $\mathcal{A}_i = \{100, 300, 500\}$) or continuous (e.g., $\mathcal{A}_i = [0, 1000]$); an individual attribute configuration is denoted $\alpha = (\alpha_1, \ldots, \alpha_n)$, where $\alpha_i \in \mathcal{A}_i$. Each observation $o_i = \langle s_i, a_i \rangle$ is a state-action pair capturing a single step of the agent's behaviour. Finally, $Prob$ represents the recogniser's belief over attribute space $\mathcal{A}$, updated as observations are received.

Recent advances in GR have incorporated Reinforcement Learning (RL) to improve adaptability in stochastic or partially observable environments [Amado *et al.*, 2022]. This adaptability creates an opportunity to extend RL-based frameworks beyond goal inference toward the recognition of broader agent attributes. Building on this line of work, we adopt and generalise the recent RL-based AR framework [Mantik *et al.*, 2025].

The RL-based AR framework in [Mantik *et al.*, 2025] consists of three main stages: (i) modelling agent attributes, (ii) offline training, and (iii) online inference.

2.1 Modelling Agent Attributes

In RL-based frameworks, the task is framed as an MDP $\mathcal{M} = \langle S, A, T, \gamma, R \rangle$, with states S, actions A, transition function T, discount factor γ, and reward function R. While RL-based GR first applied MDPs to model goals [Amado *et al.*, 2022], attributes in RL-based AR can be modelled through any MDP component, provided their influence on behaviour is expressed through agent-environment interaction [Mantik *et al.*, 2025].

We consider two attribute types following the specification in [Mantik *et al.*, 2025]. Preferences, encoded in R as per-object rewards weighted by the agent's preference value α_i; and ability levels, encoded in T as the probability of successfully transitioning to a goal state upon a skilled action, which varies jointly with the agent's ability α_i and item difficulty. Thus, agents with identical actions but differing attributes may produce distinct behavioural trajectories.

2.2 Offline Training

The offline training stage involves learning policies for the domain model via RL. Specifically, the RL problem is to learn a policy $\pi(a|s)$ for an MDP, which defines the probability of selecting action $a \in A$ in state $s \in S$. A policy π can be derived from a Q-function $Q(s, a)$, which estimates the expected return of executing action a in state s and following the policy thereafter. In the context of AR, this can be viewed as identifying which Q-function $Q(s, a)$ best explains the observed behaviour O:

$$\pi(a \mid s) = Q(s, a) / \sum_{a' \in A} Q(s, a') \tag{1}$$

The Q-function is typically learned using standard RL algorithms such as Deep Q-Networks (DQN) [Mnih *et al.*, 2015], which iteratively update Q-values through the Bellman equation until convergence. The resulting Q-function encodes how different actions are valued in each state, effectively capturing the agent's decision-making strategy.

In [Mantik *et al.*, 2025], the Q-function is augmented to include an attribute configuration α, yielding $Q_\alpha(s, a)$. While α is formally defined as a configuration that could contain multiple attribute types, their experiments focus on configurations containing only a single attribute type $\mathcal{A}_i$ and enables fine-grained inference along that attribute. We extend the Q-function $Q_\alpha(s, a)$ to handle multiple attribute types in Section 3.1.

2.3 Online Inference

With the learned Q-functions, the framework infers agent attributes from observed trajectories during online execution. For each step i in observation O, the recogniser captures a state–action pair $o_i = \langle s_i, a_i \rangle$ and compares it against the learned Q-functions $Q_\alpha(s_i, a_i)$. This comparison serves as a distance measure that quantifies how well each attribute configuration explains the observed behaviour.

Mantik *et al.* employs Bayesian Inference (BI) as the distance measure, enabling direct estimation of a probability distribution over attributes while incorporating prior knowledge about their likelihoods. Formally, the posterior probability of an attribute configuration α given an observation $o_i \in O$ is defined as:

$$P(\alpha \mid o_i) = \frac{P(o_i \mid \alpha)\, P(\alpha)}{P(o_i)}. \tag{2}$$

Here, $P(o_i \mid \alpha)$ represents the likelihood that observation o_i would occur given attribute configuration α. In the RL-based AR, this likelihood is computed from the learned Q-function: a higher Q-value $Q_\alpha(s_i, a_i)$ for the observed action a_i indicates that attribute configuration α makes that action more likely, thus increasing $P(o_i \mid \alpha)$. Specifically, the likelihood is derived from the policy in Equation 1, treating the observed action probability $\pi_\alpha(a_i \mid s_i)$ as the likelihood term.

This probabilistic formulation allows the recogniser to update its belief about the agent's underlying attributes as new observations are received, providing a principled mechanism for online inference. However, a key limitation of prior implementations is that they treat α as a discrete variable, which restricts the framework's ability to capture continuous variations in agent attributes. We address this limitation in Section 3.2, where we extend the online inference stage to support continuous-valued attribute estimation.

3 Multivariate Continuous AR Framework

This section presents the multivariate continuous AR framework, which extends the discrete RL-based AR framework described in Section 2. The framework preserves the same three-stage structure as the original formulation, while introducing: (a) a multivariate inference component, which extends the offline training process by learning a single Q-function that jointly represents multiple attribute types, and (b) a continuous inference component, which augments the online inference stage by fitting smooth probability density functions to approximate continuous posterior distributions over attributes.

3.1 Extending to Multivariate Inference

In prior works, RL policies were trained separately for each attribute type as highlighted in Section 2.2. This becomes increasingly intractable as the number of attributes grows. When attributes are combined, the number of possible attribute configurations increases combinatorially, making it impractical to train and store a separate policy for every combination. This motivates the need for a single function that jointly represents multiple attributes.

To address this, we extend the Q-function introduced in the previous framework to generalise across multiple attributes simultaneously. Instead of conditioning on a single attribute type $\alpha_i \in \mathcal{A}_i$, the Q-function is modified to take as input the full attribute configuration $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$, where each α_i corresponds to one attribute type. The resulting function $\hat{Q}_\alpha(s, a)$ represents the expected future rewards for taking action a in state s, given the combined attribute configuration α.

To condition the Q-function on multiple attributes, the offline training stage is modified by associating one attribute configuration α to each episode. At the beginning of each training episode, a set of attribute values α_i is sampled from the corresponding attribute range $\mathcal{A}_i$, forming a complete attribute configuration $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$. Each sampled attribute configuration is incorporated into the state representation, effectively treating attributes as part of the environment context during learning. Over many training episodes, the Q-function is exposed to a wide variety of attribute combinations. As a result, the learned function $\hat{Q}_\alpha(s, a)$ develops the capacity to produce distinct behaviours conditioned on different attribute configurations, enabling it to generalise across the entire attribute space $\mathcal{A}$.

3.2 Handling Continuous Values

While previous sections described inference over discrete attribute values, many agent attributes vary on a continuous scale. To capture these variations, our approach extends BI introduced in Section 2.3 to operate over continuous values. Specifically, we approximate a continuous posterior density by wrapping the BI function within a four-step pipeline that estimates the joint posterior distribution over attribute configuration $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$:

1. Draw a fixed number of samples $\{\alpha^j\}_{j=1}^m$ from the joint attribute space $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}2 \times \cdots \times \mathcal{A}n$.
2. Assign initial unnormalised *prior* density values (typically uniform unless domain-specific knowledge provides a more informative prior).
3. Compute unnormalised *posterior* density values y^j at each sampled point based on the posterior probabilities obtained from Equation 2.
4. Fit a scaled continuous probability density function (pdf) $C\text{pdf}_\theta(\alpha)$ to the dataset $\{\alpha^j, y^j\}_{j=1}^m$.

A key subtlety is that each y^j produced by BI from Equation 2 is unnormalised posterior density of a sampled attribute configuration α^j. Unlike a discrete probability distribution where all probability masses sum to one, a density value at a single point carries no such constraint, only the area under the full density curve must equal one. Consequently, when fitting a continuous pdf we must jointly estimate both the distribution parameters θ and a global scaling constant C:

$$y^j = C\,\text{pdf}_\theta(\alpha^j), \tag{3}$$

Attribute-Specific Transformations
Before fitting, we apply a transformation to each attribute that maps it into a space where a distribution is appropriate. Our formulation accommodates attributes with differing statistical properties by allowing separate assumptions for their underlying continuous distributions. These assumptions determine how the unnormalised posterior densities y^j obtained from BI are mapped and fitted before being combined in the multivariate setting. Regardless, even if the correct distributional shape is unknown, one can always fall back to a simple uniform distribution on a bounded interval.

In this study, we focus on two attribute types—ability level and preference—but the proposed procedure can be readily extended to additional attributes, provided that appropriate distributional assumptions are specified.

Ability Level Attribute. The ability level attribute $\alpha_{ab} \in \mathcal{A}_{ab} = [L, U]$ is a bounded continuous variable capturing the agent's competence level. We apply a logarithmic transformation $\alpha'_{ab} = \log \alpha_{ab}$ and fit a truncated log-normal distribution, which captures potential asymmetry while preserving the bounded support $[L, U]$. In contrast to the Normal distribution, which is strictly symmetric, the log-normal can capture asymmetry, making it more appropriate for modelling the observed characteristics of the ability level attribute.

Preference Attribute. The preference attribute $\alpha_{pr} = (p_1, p_2, \ldots, p_k) \in \mathcal{A}_{pr}$ represents the relative importance or desirability an agent assigns to different possible outcomes, where each $p_j \geq 0$ and $\sum_{j=1}^k p_j = 1$, placing α_{pr} on a $(k-1)$-simplex Δ^{k-1}.

Because the sum-to-one constraint reduces the effective dimensionality, no standard multivariate distribution (including the Normal distribution) can be fitted directly. The Additive Log-Ratio (ALR) transformation resolves this by expressing each preference as a log-ratio relative to a reference category p_k, mapping the simplex to unconstrained $\mathbb{R}^{k-1}$:

$$\alpha'_{pr,j} = \log \frac{p_j}{p_k}, \qquad j = 1, 2, \ldots, k-1, \quad \alpha'_{pr} \in \mathbb{R}^{k-1}. \tag{4}$$

This transformation effectively removes the simplex constraint while preserving the relative ratios among preferences, after which a standard Normal distribution can be fitted.

Multivariate Fitting

After defining the attribute-specific transformations, we jointly fit a continuous pdf that captures potential correlations between attributes.

For the attribute space of ability level and preference $\mathcal{A} = \mathcal{A}_{ab} \times \mathcal{A}_{pr}$, we define the transformed variable as:

$$z = [\log \alpha_{ab},\ \alpha'_{pr}] \in \mathbb{R}^k, \tag{5}$$

where $\log \alpha_{ab}$ represents the log-transformed ability and α'_{pr} denotes the ALR-mapped preference vector. We assume that z follows a multivariate truncated Normal distribution, $z \sim \mathcal{N}(\mu, \Sigma)$.

Fitting the Normal distribution in the transformed space is straightforward, but the unnormalised density values y^j were computed in the original space $(\alpha_{ab}, \alpha_{pr})$. We therefore need a formula that translates the fitted $pdf_\theta(\alpha_{ab}, \alpha_{pr})$ back into the original space.

The Change of Variable Theorem provides this correction by multiplying the transformed density by the absolute Jacobian determinant of the transformation, ensuring $pdf_\theta(\alpha_{ab}, \alpha_{pr})$ integrates correctly in the original space, such that:

$$pdf_\theta(\alpha_{ab}, \alpha_{pr}) = \mathcal{N}([\log \alpha_{ab},\ \alpha'_{pr}] \mid \mu, \Sigma) \left| \frac{d([\log \alpha_{ab},\ \alpha'_{pr}])}{d[\alpha_{ab},\ \alpha_{pr}]} \right|, \tag{6}$$

where $\frac{dz}{d\alpha}$ is the Jacobian matrix of partial derivatives describing how the transformation $z = [\log \alpha_{ab},\ \alpha'_{pr}]$ depends on the original attributes $\alpha = [\alpha_{ab},\ \alpha_{pr}]$. Note that we can further simplify the Jacobian determinant,

$$\left| \frac{d([\log \alpha_{ab},\ \alpha'_{pr}])}{d[\alpha_{ab},\ \alpha_{pr}]} \right| = \frac{1}{\alpha_{ab} \prod_{j=1}^{k} p_j}. \tag{7}$$

Furthermore, a standard Normal assigns probability mass across the entire real line, including outside $[L, U]$. Since the ability variable α_{ab} is bounded within $[L, U]$, we introduce a truncation normalisation term to ensure the pdf integrates to one within the valid range,

$$\frac{1}{\Phi\left(\frac{\log U - \mu_{[0]}}{\sqrt{\Sigma_{[00]}}}\right) - \Phi\left(\frac{\log L - \mu_{[0]}}{\sqrt{\Sigma_{[00]}}}\right)}. \tag{8}$$

Combining the Change of Variable correction and the truncation term, the joint density in the original attribute space is expressed as:

$$pdf_\theta(\alpha_{ab}, \alpha_{pr}) = \mathcal{N}([\log \alpha_{ab},\ \alpha'_{pr}] \mid \mu, \Sigma) \frac{1}{\alpha_{ab} \prod_{j=1}^{k} p_j} \frac{1}{\Phi\left(\frac{\log U - \mu_{[0]}}{\sqrt{\Sigma_{[00]}}}\right) - \Phi\left(\frac{\log L - \mu_{[0]}}{\sqrt{\Sigma_{[00]}}}\right)}. \tag{9}$$

Item	Ingredients	Craft Location	Craft Difficulty
Cloth	Grass	Factory	100
Stick	Wood	Workbench	100
Plank	Wood	Toolshed	500
Rope	Grass	Toolshed	500

Table 1: Item recipes and associated difficulties in Craft World.

To fit this distribution to the unnormalised posterior densities (α^j, y^j), we jointly estimate the parameters $\theta = \{\mu, \Sigma\}$ and the global scaling constant C by minimising:

$$\hat{C}, \hat{\theta} = \arg\min_{C,\theta} \sum_j \left(y^j - C\, pdf_\theta(\alpha^j_{ab}, \alpha^j_{pr})\right)^2. \tag{10}$$

Once the continuous pdf has been fitted, inference over attributes is performed by evaluating the posterior density across the attribute space. Given new behavioural observations, posterior likelihoods are computed using BI, and the fitted scaled pdf $C\, pdf_{\hat{\theta}}(\alpha_{ab}, \alpha_{pr})$ is used to obtain continuous estimates of the most probable attribute values. The most likely attribute configuration is then obtained as:

$$\hat{\alpha} = \arg\max_{\alpha \in \mathcal{A}} C\, pdf_{\hat{\theta}}(\alpha), \tag{11}$$

where $\alpha = (\alpha_{ab}, \alpha_{pr})$. This formulation yields a continuous and differentiable posterior density that supports joint reasoning over the entire attribute space. In practice, we generate a large set of candidate attribute values, evaluate pdf at each candidate, and return the one with the highest posterior value.

4 Experiments

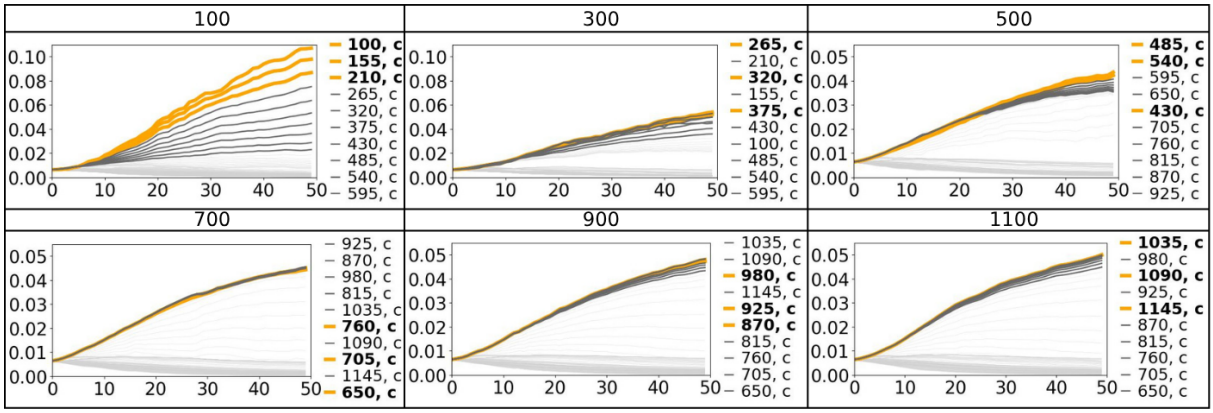
We evaluate our framework in Craft World [Dann *et al.*, 2023], a grid-based crafting environment where agents collect materials and craft items. We introduce a new scenario for evaluating continuous preference and ability inference. Agents craft four items (cloth, stick, plank, rope) with varying difficulty (Table 1). Items sharing ingredients or locations create behavioural ambiguity, making it insufficient for the observer to infer attributes purely from which materials are collected. Furthermore, higher-ability agents can exploit overlapping crafting locations to produce multiple items, enabling richer behavioural variation.

We evaluate the framework through two experiments: (1) multivariate discrete inference, and (2) multivariate continuous inference.[1] All policies are trained using DQN [Mnih *et al.*, 2015], and each experiment comprises 100 randomised episodes of 50 steps. The observer and the acting agent are trained using separate Q-functions to assess the framework's ability to generalise across independently trained models.

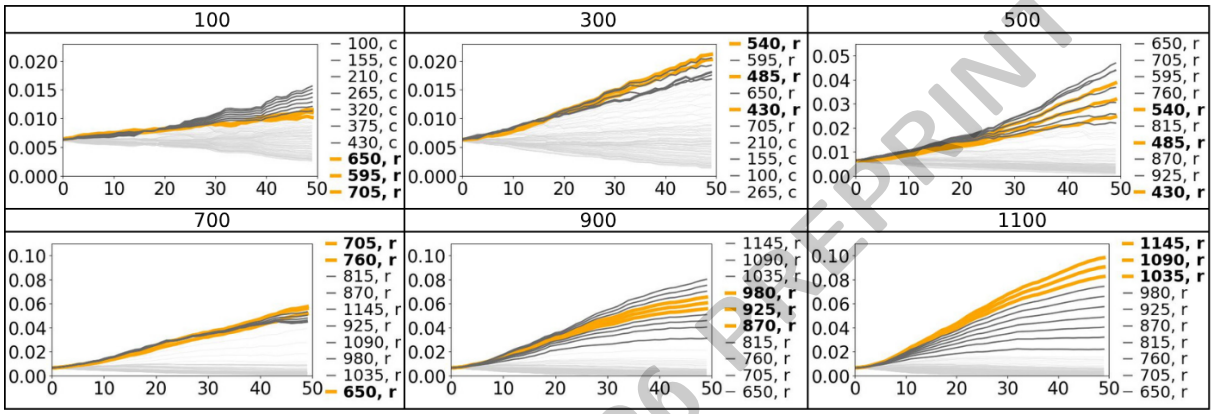
4.1 Multivariate Discrete Inference

This experiment evaluates the framework's ability to jointly infer discrete-valued attributes from observations. It serves as a discrete baseline to verify that the multivariate inference component introduced in Section 3.1 functions as intended.

[1] See: https://github.com/sh-ryl/IJCAI26_ContinuousMultiAR



(a) Cloth-preferring agent



(b) Rope-preferring agent

Figure 1: Multivariate **discrete** inference results. Each figure corresponds to a ground-truth preference, with cells showing inferred posteriors for each ability level (cell title, arranged across two rows). Lines show posterior probabilities over episode timesteps (x-axis), with the y-axis reflecting inference confidence. The legend lists the top-10 configurations out of 160 total, ranked by average posterior across all timesteps. Thick orange lines highlight the three configurations closest to ground truth; dark grey lines indicate the remaining top-10 configurations; light grey lines represent all others.

For the discrete case, posterior likelihoods are obtained directly from the BI formulation (Equation 2), comparing observed trajectories with the attribute-conditioned Q-function before extending to the continuous setting.

Experiment Setup

The observer is instantiated using the discrete attribute configurations defined below, and the attribute space $\mathcal{A} = \mathcal{A}_{ab} \times \mathcal{A}_{pr}$ is assumed to be confined to predefined combinations of ability level and preference values.

The agent's preference ($\mathcal{A}_{pr}$) is represented as a four-dimensional vector over the craftable items cloth, stick, plank, and rope, where each element denotes the weight assigned to crafting that item, and all values sum to one. Two preference strengths are considered: strong preferences with one dominant value of 0.7 and three lower values of 0.1, and weak preferences with one dominant value of 0.4 and three lower values of 0.2. This yields eight configurations: (0.7, 0.1, 0.1, 0.1), (0.1, 0.7, 0.1, 0.1), (0.1, 0.1, 0.7, 0.1), (0.1, 0.1, 0.1, 0.7) for strong preferences, and (0.4, 0.2, 0.2, 0.2), (0.2, 0.4, 0.2, 0.2), (0.2, 0.2, 0.4, 0.2), (0.2, 0.2, 0.2, 0.4) for weak preferences. The agent's ability level ($\mathcal{A}_{ab}$) is defined as discrete values ranging from 100 to 1145 in increments of 55, yielding 20 fixed points: $\{100, 155, 210, \ldots, 1145\}$. Combining both attributes yields 160 unique agent configurations, each eliciting distinct behaviour from the single attribute-conditioned policy.

We choose the following values for ground truth. Two types of preference: cloth (0.7,0.1,0.1,0.1) and rope (0.1,0.1,0.1,0.7), both requiring grass but differing in difficulty. Stick and plank are omitted as they mirror cloth and rope behaviours. Ability levels span in the range of $[0, 1200]$ with six values $100, 300, 500, 700, 900, 1100$ representing low to high competence.

Results and Interpretation

Figure 1 visualises the posterior probabilities (Equation 2) for every combination of ground-truth preference and ability level, summarising how inference confidence is distributed across attribute configurations in each setting.

Ability level inference. For *cloth-preferring* agents, ability inference is accurate, with top predictions near ground truth

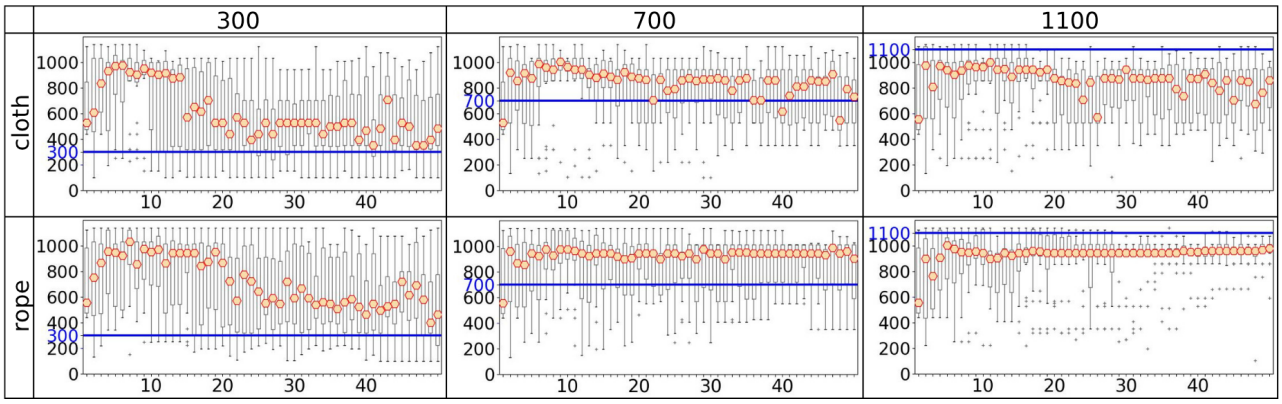


Figure 2: Multivariate **continuous ability level** inference results. Each box plot shows the distribution of maximum posterior density values per timestep across episodes, for a ground-truth ability level (columns) and preference (row) combination. Red hexagon median marker: median inferred value; box/whiskers: episode variability - a narrower spread indicates more stable inference, while a wider spread reflects greater sensitivity to episode-specific behavioural variation; blue line: ground truth attribute value.

(e.g., ground truth 100 and top-3 inference 155, 100, 210), except for ground truth 700 which predicts value around 900. This suggests that there is limited behavioural variability between 700 and 900. For *rope-preferring* agents, mid-range abilities (300–500) show lower accuracy, with predictions biased around 595-650. For 900 and 1100, inferred values saturate near 1145, as behavioural differences diminish once difficult tasks are achieved. Despite this, the framework still captures clear distinctions between low- and high-ability agents, even if intermediate levels are estimated less precisely.

Preference inference. For *cloth-preferring*, preferences are accurate for all ground truth ability levels. For *rope-preferring* agents, the top-10 inferred configurations converge more consistently toward rope as ability level increases. The inference is accurate except for one case: rope preference (0.1,0.1,0.1,0.7) at ability 100 is misclassified as cloth (0.7,0.1,0.1,0.1). This happens because both items share the same ingredient (grass), and low-ability agents fail to craft. This results in low inference confidence (all configurations below 2%), suggesting uncertainty over misclassification.

Overall, inference confidence correlates with task difficulty: easier items (cloth) show higher confidence for lower abilities, while difficult items (rope) show higher confidence for higher abilities. The discrete multivariate framework successfully infers preference and ability level jointly while capturing meaningful behavioural trends, with variability reflecting observability limits rather than model inconsistency.

4.2 Multivariate Continuous Inference

While discrete inference captures attribute relationships at fixed sampled points, real-world attributes often vary along continuous scales. This experiment evaluates whether the framework can infer such attributes continuously and in combination by extending the multivariate inference formulation to continuous values. The observer follows the continuous inference pipeline introduced in Section 3.2 to approximate the joint posterior density over preference and ability level as a smooth function.

Experiment Setup

The observer is instantiated using the same configurations defined in Section 4.1 with an additional assumption that the attribute space $\mathcal{A}$ spans continuous ranges for both preference and ability level. For preference $\mathcal{A}_{pr}$, each of the four items can take any value between 0 and 1, subject to the constraint that all components sum to one. For ability level $\mathcal{A}_{ab}$, the range is defined over $[0, 1200]$.

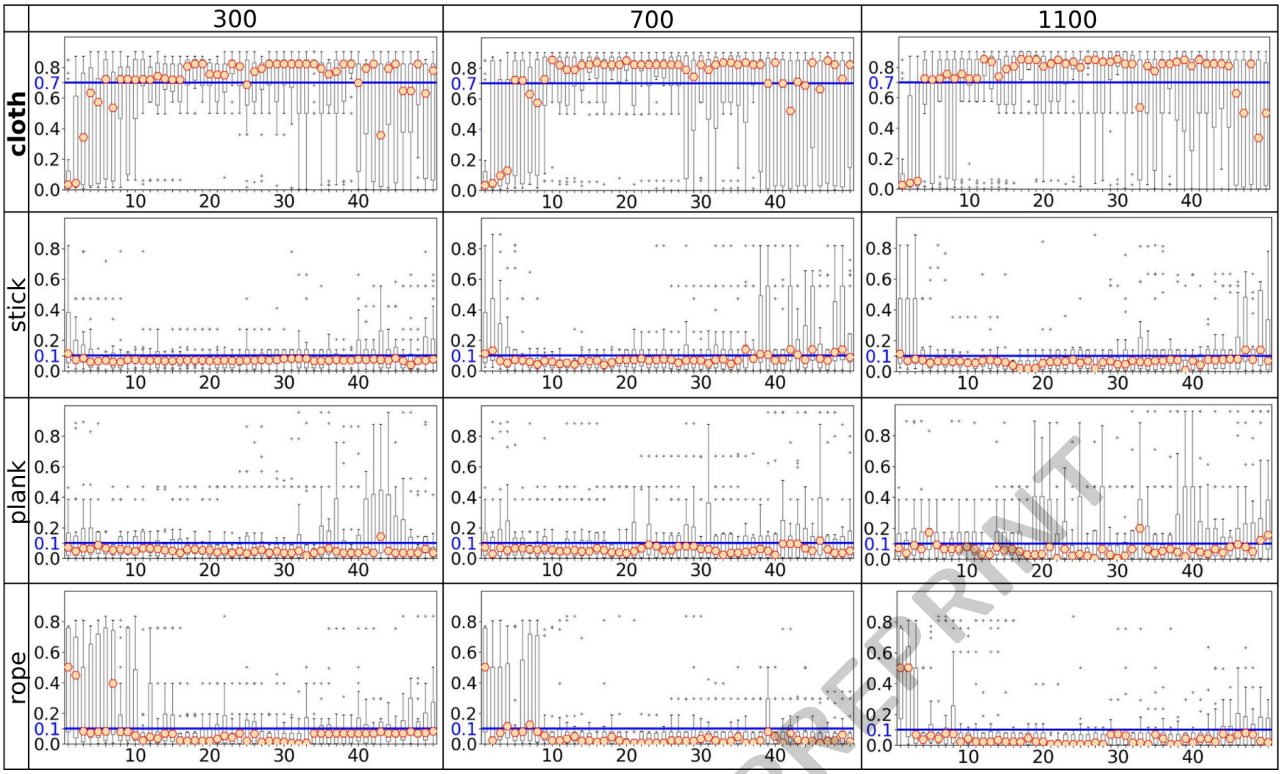
In the first step of the pipeline described in Section 3.2, we draw fixed samples across the attribute space. The values used for sampling correspond to the same discrete attribute configurations defined in Section 4.1.

Results and Interpretation

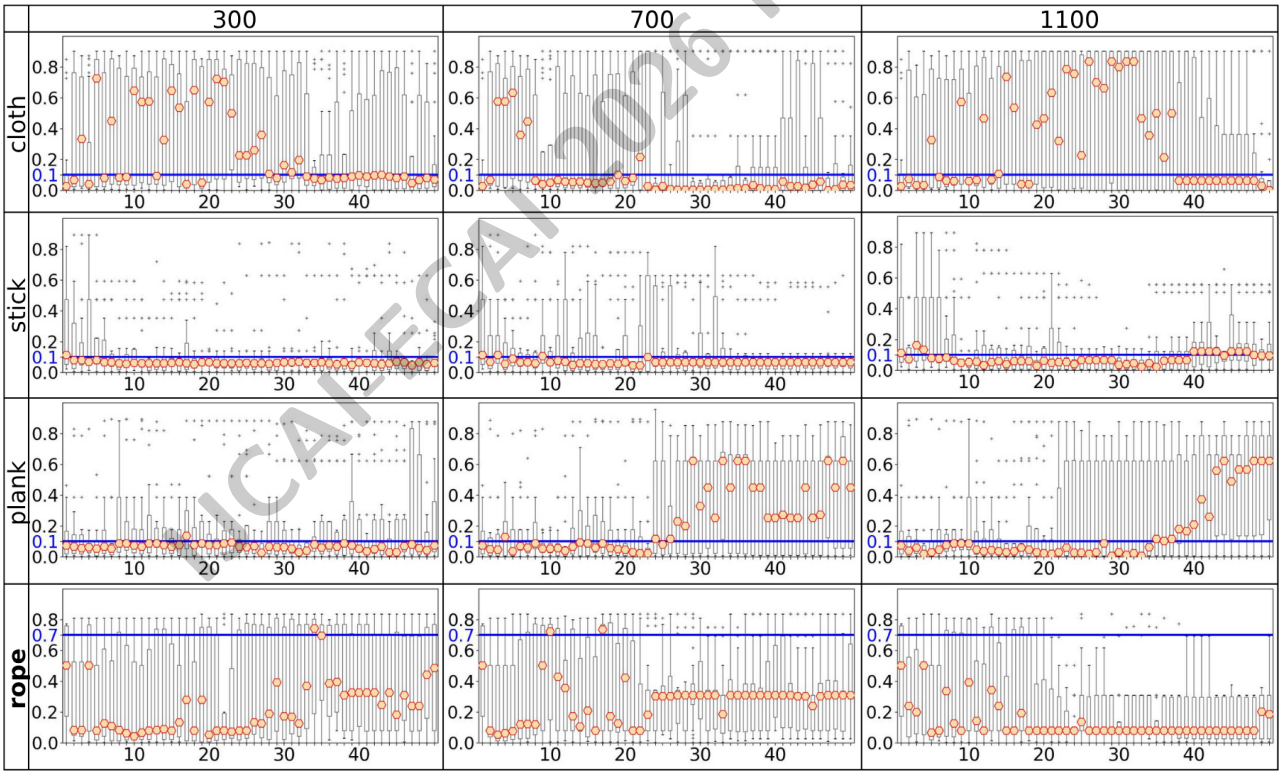
Figure 2 and 3 visualise the evolution of the observer's inference over time for both ability level and preference. To aid readability, we chose three representative ground truth for ability level, 300 (low), 700 (moderate), and 1100 (high), with preference of cloth (0.7,0.1,0.1,0.1) and rope (0.1,0.1,0.1,0.7).

Ability level inference. In Figure 2, inference converges within approximately 20 timesteps. Across both preference types, ability 300 converges toward 500 while 700 and 1100 center near 900, with spread narrowing for higher abilities (400–1000 for 300 vs 600–1000 for 700 and above). The exception is rope-preferring agents at ability 1100, where the interquartile range narrows further (800–1000), suggesting more consistent inference when high ability aligns with a difficult task.

Preference inference. In Figure 3, the inferred preference varies across different stages of the episode. For *cloth-preferring* agents, inferred cloth preference remains high (0.8) while stick/plank/rope stay low (0.1). After 30 timesteps, cloth variability widens slightly, likely reflecting behavioural inconsistency as ingredients deplete. For *rope-preferring* agents, inferred rope preference is generally underestimated at low ability (300), converging to 0.4 rather than the true value of 0.7. At higher ability levels (700–1100), inference exhibits early fluctuation before stabilising,



(a) Cloth-preferring agent



(b) Rope-preferring agent

Figure 3: Multivariate **continuous preference** inference results. Each box plot shows the distribution of maximum posterior density values per timestep across episodes, for a ground-truth ability level and preference configuration (columns) and an inferred attribute value (rows). See Figure 2 for plot element descriptions.

though convergence accuracy varies across levels. Plank median stays near 0.1 at ability level 300, but gradually increases toward the end of the episode for abilities 700 and 1100, with the increase more pronounced at higher abilities. This pattern for both plank and rope suggests that higher-ability agents craft rope earlier in the episode, and once the ingredient for rope is depleted, higher-ability agents begin crafting additional items such as plank. Cloth and stick preferences remain low (0.1), though the high variability for cloth was caused by cloth and rope sharing grass as an ingredient (Table 1).

These results indicate that the observer dynamically adapts its inference as behaviour evolves within an episode, a property that emerges naturally from the sequential Bayesian update mechanism rather than being explicitly designed in. When agents switch from rope to other items after resource depletion, preference inference updates accordingly, showing the framework continually re-evaluates attribute configurations rather than assuming fixed attributes. This sequential updating produces a moving-average effect that weights recent observations more strongly, achieving sensitivity to evolving behaviour without requiring explicit temporal smoothing parameters, similar to [Dann *et al.*, 2023].

Overall, the continuous inference results show that the framework achieves stable ability-level inference and adjusts preference inference over time, reflecting its capacity to model continuous agent attributes under varied conditions.

5 Related Work

Attribute Recognition (AR) shares its motivation with agent modelling, which seeks to reason about other agents by inferring the internal factors that drive their behaviour [Albrecht and Stone, 2018; Van-Horenbeke and Peer, 2021]. Across its formulations, the key challenge lies in representing and inferring these latent factors, referred to here as attributes. While terminology varies across subfields, most studies agree that goals, preferences, and beliefs are fundamental internal dynamics shaping agent behaviour [Georgeff and Rao, 1991; Baker *et al.*, 2017; Ramírez and Geffner, 2010].

A foundational line of work is Bayesian inverse planning, which models behaviour as probabilistic inference over latent mental variables. It infers goals, preferences, or beliefs by inverting a forward planning model and estimating their posterior distributions from observed trajectories [Baker *et al.*, 2009; Baker *et al.*, 2017; Evans *et al.*, 2015]. These approaches combine forward planning with Bayesian updating to interpret behaviour as evidence for underlying mental states, typically over discrete attribute spaces. Our framework adopts this probabilistic perspective to continuous attribute representations, enabling fine-grained reasoning over behavioural variation.

Another major formulation, Inverse Reinforcement Learning (IRL), seeks to recover an agent's reward function that best explains its observed actions [Ng and Russell, 2000; Lee *et al.*, 2017]. IRL and its variants have been extended to handle partial observability, multi-task settings, and non-stationary reward structures [Arora and Doshi, 2018]. However, IRL reduces all behavioural factors to a single reward signal, potentially obscuring distinct influences on decision-making. We instead model multiple interpretable continuous attributes that jointly explain behaviour without collapsing them into a single objective.

Our technical approach relates to several RL paradigms. Multi-objective Q-learning learns policies balancing multiple reward signals [Hayes *et al.*, 2021]. Like our approach, it represents multiple factors in the Q-function, but these factors are known objectives the agent optimises, not latent attributes to be inferred from observations. Multi-agent methods like QMIX [Rashid *et al.*, 2018] and REFIL [Iqbal *et al.*, 2020] condition policies on agent-specific features for coordination, sharing our use of conditioning mechanisms but focusing on learning behaviour rather than inferring agent attributes.

More recently, research in computational Theory of Mind (ToM) extends these ideas by aiming to infer and reason about agents' mental states [Rabinowitz *et al.*, 2018; Jara-Ettinger, 2019; Wu *et al.*, 2023]. ToM-inspired models learn to predict others' intentions or future actions based on latent representations of beliefs and goals. These models demonstrate strong generalisation across agents and tasks but often prioritise predictive performance over interpretability. Our framework complements this direction by modelling continuous latent attributes that are explicitly defined and jointly inferred.

Overall, prior research shares the goal of inferring latent attributes that jointly explain behaviour. Building on this foundation, we introduce a unified probabilistic formulation of Attribute Recognition that enables continuous and interpretable joint inference over multiple latent attributes. This formulation extends existing paradigms by capturing fine-grained behavioural variation while maintaining conceptual clarity and interpretability across attributes.

6 Conclusion

We presented an RL-based Attribute Recognition framework that enables joint inference over multiple continuous-valued agent attributes by conditioning a single Q-function on multiple attributes and performing Bayesian inference through continuous probability density fitting. To the best of our knowledge, this is the first approach in both Goal Recognition and Attribute Recognition to infer continuous attribute values directly from behaviour. Through experiments in the Craft World domain, we showed that the framework reliably produces stable and fine-grained inference across continuous attribute spaces while avoiding the need to train separate policies for each attribute configuration. In particular, the observer rapidly converges to ability level estimates close to the ground truth and dynamically refines preference estimates as behaviour unfolds, even under behavioural ambiguity and resource constraints. These results demonstrate that modelling agent attributes as continuous variables within a unified inference model improves scalability, expressiveness, and robustness compared to discrete alternatives.

References

[Albrecht and Stone, 2018] Stefano V. Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95, 2018.

[Amado *et al.*, 2022] Leonardo Amado, Reuth Mirsky, and Felipe Meneguzzi. Goal recognition as reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(9):9644–9651, Jun. 2022.

[Arora and Doshi, 2018] Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artif. Intell.*, 297:103500, 2018.

[Baker *et al.*, 2009] Chris L. Baker, Rebecca Saxe, and Joshua B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 113:329–349, 2009.

[Baker *et al.*, 2017] Chris L. Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B. Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1, 2017.

[Dann *et al.*, 2023] Michael Dann, Yuan Yao, Natasha Alechina, Brian Logan, Felipe Meneguzzi, and John Thangarajah. Multi-agent intention recognition and progression. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 91–99. International Joint Conferences on Artificial Intelligence Organization, 8 2023. Main Track.

[Evans *et al.*, 2015] Owain Evans, Andreas Stuhlmüller, and Noah D. Goodman. Learning the preferences of ignorant, inconsistent agents. In *AAAI Conference on Artificial Intelligence*, 2015.

[Georgeff and Rao, 1991] M Georgeff and A Rao. Modeling rational agents within a bdi-architecture. In *Proc. 2nd Int. Conf. on Knowledge Representation and Reasoning (KR'91). Morgan Kaufmann*, pages 473–484. of, 1991.

[Hayes *et al.*, 2021] Conor F. Hayes, Roxana Ruadulescu, Eugenio Bargiacchi, Johan Kallstrom, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai Aravazhi Irissappane, Patrick Mannion, Ann Now'e, Gabriel de Oliveira Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36, 2021.

[Iqbal *et al.*, 2020] Shariq Iqbal, C. S. D. Witt, Bei Peng, Wendelin Bohmer, Shimon Whiteson, and Fei Sha. Randomized entity-wise factorization for multi-agent reinforcement learning. In *International Conference on Machine Learning*, 2020.

[Jara-Ettinger, 2019] Julian Jara-Ettinger. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29:105–110, 2019.

[Kautz and Allen, 1986] Henry A. Kautz and James F. Allen. Generalized plan recognition. In *Proceedings of the Fifth AAAI National Conference on Artificial Intelligence*, AAAI'86, page 32–37. AAAI Press, 1986.

[Lee *et al.*, 2017] Kamwoo Lee, Mark Rucker, William T. Scherer, Peter A. Beling, Matthew S. Gerber, and Hyojung Kang. Agent-based model construction using inverse reinforcement learning. *2017 Winter Simulation Conference (WSC)*, pages 1264–1275, 2017.

[Mantik *et al.*, 2025] Sheryl Mantik, Michael Dann, Minyi Li, Huong Ha, and Julie Porteous. Beyond goal recognition: A reinforcement learning-based approach to inferring agent behaviour. In *Adaptive Agents and Multi-Agent Systems*, 2025.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Kirkeby Fidjeland, Georg Ostrovski, Stig Petersen, Charlie Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

[Ng and Russell, 2000] Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, 2000.

[Rabinowitz *et al.*, 2018] Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew M. Botvinick. Machine theory of mind. *ArXiv*, abs/1802.07740, 2018.

[Ramírez and Geffner, 2010] Miquel Ramírez and Hector Geffner. Probabilistic plan recognition using off-the-shelf classical planners. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'10, page 1121–1126. AAAI Press, 2010.

[Rashid *et al.*, 2018] Tabish Rashid, Mikayel Samvelyan, C. S. D. Witt, Gregory Farquhar, Jakob N. Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. *ArXiv*, abs/1803.11485, 2018.

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

[Wu *et al.*, 2023] Haochen Wu, Pedro Sequeira, and David V. Pynadath. Multiagent inverse reinforcement learning via theory of mind reasoning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, page 708–716, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems.