Technical Supplement

1 Proofs of Lemmas and Theorems

Lemma 1. *If an* MDP *has a proper policy π, then any policy which is ϵ-greedy with respect to A is also proper.*

Outline Proof. Lemma 1 follows directly from the Definition 1 of a proper policy: If there exists some proper policy π, then in all states s, any ϵ-greedy policy has a non-zero probability of executing action $a = \pi(s)$. Since the ϵ-greedy policy has a non-zero chance of mimicking proper policy π in every state, it is also proper. □

Lemma 2. *Consider an* FMDP *where π_+ is proper, an agent with awareness $\mathcal{X}^t \subseteq \mathcal{X}^+, A^t \subset A^+$, and expert acting with respect to (10-13). If $\exists a \in image(\pi_+), a \notin A^t$ then as $k \to \infty$, either $Err(t, t+k) \to c$ with $c \leq \beta$ or the expert utters (12) such that $a' \notin A^t$.*

Outline Proof. If $Err(t, t+k) \to c \leq \beta$, we are done. If not, we must consider two cases—one where an ϵ-greedy policy over A^t is proper, and one where it is not. If it is, (11) will eventually be satisfied, since the expert will gather enough samples to approximate the agent's policy error (which is above β). Further, we know $\exists s, \pi_+(s) = a' \wedge a' \notin A^t$, and that s is reachable by the agent, so (12) will eventually be satisfied. If the agent's policy is *not* proper, the agent may visit some set of states $\mathcal{S}' \subset \mathcal{S} \backslash \mathcal{S}_e$ infinitely often (thus satisfying (11) for finite κ). Since π_+ is proper, there must exist some $s \in \mathcal{S}'$ such that $\exists a' \notin A^t, \forall a \in A^t, Q_{\pi_+}(s, a') > Q_{\pi_+}(s, a)$, thus satisfying (12). In either case, (10) will eventually be satisfied for finite μ. Since (10-12) are not mutually exclusive, all three will eventually be true simultaneously, causing the expert to utter (13). □

Lemma 3. *Consider an* FMDP *where π_+ is proper and an agent with awareness $\mathcal{X}^t \subset \mathcal{X}^+, image(\pi_+) \subseteq A^t \subseteq A^+$. If $\exists s \exists s' \neq s, s[\mathcal{X}^t] = s'[\mathcal{X}^t]$, and $\pi_+(s) \neq \pi_+(s')$, then as $k \to \infty$, either $Err(t, t+k) \to c$ $(c \leq \beta)$, or the expert utters (20) such that $X \notin \mathcal{X}^t$*

Outline Proof. The ϵ-greedy agent is aware of all actions in $image(\pi_+)$, so can visit all states reachable via π_+, including s and s'. If $Err(t, t+k) \to c, c \leq \beta$, we are done. If not, then at some time $t + k_1$ the agent will do a' in s, causing the expert to utter $Q(w^s_{t+k_1}, a) > Q(w^s_{t+k_1}, a')$. Similarly, at some time $t + k_2$, the agent will receive advice $Q(w^s_{t+k_2}, a') > Q(w^s_{t+k_2}, a)$. Since $s_{t+k_1}[\mathcal{X}^t] = s_{t+k_2}[\mathcal{X}^t]$, the two pieces of advice appear to the agent to conflict, so the agent will ask (19) with answer $X \notin \mathcal{X}^t$. □

Lemma 4. *Consider an* FMDP *where π_+ is proper and an agent with awareness $A^t \subseteq A^+$, $\mathcal{X}^t \subseteq \mathcal{X}^+$, $scope_t(\mathcal{R}) \subseteq scope_+(\mathcal{R})$. As $k \to \infty$, there exists a K such that for all $k \geq K$, $\mathcal{R}_{t+k}(s) = \mathcal{R}_+(s)$ for all states s reachable using A^t.*

Outline Proof. If s is reachable using A^t, then as $k \to \infty$, an ϵ-greedy agent will eventually enter s at some time i, receive reward $\mathcal{R}_+(s)$, and update its current reward function so that $\mathcal{R}_i(s) = \mathcal{R}_+(s)$. If the agent has previously encountered another s' such that $s[scope_t(\mathcal{R})] = s'[scope_t(\mathcal{R})]$ and $\mathcal{R}_+(s) \neq \mathcal{R}_+(s')$, the partial descriptions (21) for s and s' will conflict. The agent resolves this by asking (22), receiving an answer differentiating s from s' in $\mathcal{R}_+$. □

Theorem 1. *Consider an* FMDP *where π_+ is proper and an agent with initial awareness $\mathcal{X}^0 \subseteq \mathcal{X}^+, A^0 \subseteq A^+$, and $scope_0(\mathcal{R}) \subseteq scope_+(\mathcal{R})$ acts according to algorithm 2. If for all $X \in \mathcal{X}^+$, there exists a pair of states s, s' such that $s[\mathcal{X}^+ \setminus X] = s'[\mathcal{X}^+ \setminus X]$, $s[X] \neq s'[X]$, and $\pi_+(s) \neq \pi_+(s')$, then as $t \to \infty$, $Err(0, t) \to c$ such that $c \leq \beta$*

Outline Proof. By repeatedly applying theorems 2 - 4, we have that if $Err(0, t)$ has not yet converged to $c \leq \beta$, then there exists a K where $image(\pi_+) \subseteq A^K$, $\mathcal{X}^K = \mathcal{X}^+$, and $\mathcal{R}_K(s) = \mathcal{R}_+(s)$ for all s reachable with A^K. Thus $\mathcal{X}^K$, A^K, $\mathcal{R}_K$ define a separate MDP for which the agent is fully aware, but has the same π_+ as the original. All episodes terminate, so the agent's estimate of $\mathcal{T}$ eventually approximates the true transition function, V_t converges to the V_{π_+}, and thus $Err(0,t) \to 0$. □

2 DBN structures from Coffee-Robot Experiment

Figure 1 below shows the structure of probabilistic dependencies between variables for each of the four actions in the *coffee-robot* problem. Figure 1a shows the true structure of the problem, while Figures 1b and 1c show an example of the structures learned by the *default* and *high tolerance* agents as described in Section 4 of the main paper.

The figures show that the default agent successfully learns the true structure of the decision problem. The agent paired with the high tolerance expert is also able to learn the correct dependencies, but only for the subset of variables it was made aware of by the expert.

3 Expert Messages from the Factory Experiments

This section provides additional information on the expert messages sent during the*factory* experiments in Section 4.2 of the main paper. Table 1 shows the average number of expert messages sent to the agent for each of the three settings for expert tolerance. It shows that, as expected, higher tolerances correspond to fewer messages.

Figure 2 shows the average number of expert messages sent over time. As can be seen in all three charts, the agents tend to receive the majority of expert messages early in learning. This amount tails-off towards the end of the experiment. The reason for this is that, as learning progresses, the agent learns an increasingly accurate model of the problem, and is therefore less prone to make mistakes which will be corrected by the expert, or to discover unexpected scenarios which conflict with its current understanding of the world.

4 The Language for Partially Describing Factored Markov Decision Networks

The model for learning the true FMDP (*fmdp*$_+$) in the main paper made use of a language $\mathcal{L}$ for partially describing FMDPs. This language was used in two ways: to represent monotonic information from evidence, and as a language in which the learner and expert can communicate (partial) information about the true FMDP. Its syntax and semantics are defined below.

4.1 The syntax of the language $\mathcal{L}$

Terms of the sort: $X, Y \ldots$ are *state variable* (SV) constants; Pa_Y, $\mathcal{X}$, $scope(\mathcal{R})$, are Sets of State Variables (SSV) constants (denoting sets of state variables in the model). Similarly, there are actions (A) like *move*, *buy-coffee*, *pick-up* etc. and sets of actions (SA).

An atomic state (AS) term (e.g $s_{m,n}$) denotes a full atomic state in the model. Where $\mathcal{Y}$ is an SSV and s an AS, $s[\mathcal{Y}]$ is a partial-state assignment (PS) term, denoting a value assignment to each SV constant in $\mathcal{Y}$. The language also includes SV variables and AS variables.

Additionally, there are numeric terms (N), which denote the real numbers.

A well-formed formula within the language L is then given by the following grammar:

$$\begin{aligned}\langle L\rangle ::=& \langle AS\rangle = \langle AS\rangle \mid \langle PS\rangle = \langle PS\rangle \mid \langle SV\rangle = \langle SV\rangle\\ &|\langle A\rangle \in \langle AS\rangle \mid \langle SV\rangle \in \langle SSV\rangle \mid \langle A\rangle \in \langle SA\rangle\\ &|R(\langle AS\rangle = \langle N\rangle)\\ &|Q(\langle S\rangle, \langle A\rangle) > \langle N\rangle \mid Q(\langle S\rangle, \langle A\rangle) > \langle N\rangle\\ &|\neg\langle L\rangle \mid \langle L\rangle \wedge \langle L\rangle\\ &|\exists\langle S\rangle\langle L\rangle \mid \exists\langle SV\rangle\langle L\rangle\\ &|?\lambda\langle SV\rangle\langle L\rangle\end{aligned}$$

Each model *fmdp* for interpreting $\mathcal{L}$ corresponds to a (unique) complete FMDP (see Section 2 of the main paper for a definition). Section 4.2 then evaluates the formulae of $\mathcal{L}$ as partial descriptions of *fmdp*.

4.2 The semantics of $\mathcal{L}$

Let $fmdp = \langle \mathcal{X}^{fmdp}, \mathcal{A}^{fmdp}, Pa^{fmdp}, \theta^{fmdp}, \mathcal{R}^{fmdp}\rangle$ be an FMDP and g a variable assignment function.

- For an SV constant X, $\llbracket X\rrbracket^{\langle fmdp,g\rangle} = X$; similarly for SSV, A, and SA constants.[1]
- For an AS variable s, $\llbracket s\rrbracket^{\langle fmdp,g\rangle} = g(s)$ where $g(s) \in v(\mathcal{X}^{fmdp})$. For an A variable a, $\llbracket a\rrbracket^{\langle fmdp,g\rangle} = g(a)$ where $g(a) \in \mathcal{A}^{fmdp}$. For an SV variable V, $\llbracket V\rrbracket^{\langle fmdp,g\rangle} = g(V)$ where $g(V) \in \mathcal{X}^{fmdp}$.

[1] If $X \notin \mathcal{X}^{fmdp}$ then $\llbracket X\rrbracket^{\langle fmdp,g\rangle}$ is undefined and our semantics ensures that any formula ϕ featuring X is such that $fmdp \not\models \phi$; similarly for propositional terms p featuring a value of a variable that is not a part of *fmdp*.

- For an SV term a and an SSV term b, $\llbracket a \in b \rrbracket^{\langle fmdp,g\rangle} = 1$ iff $\llbracket a \rrbracket^{\langle fmdp,g\rangle} \in \llbracket b \rrbracket^{\langle fmdp,g\rangle}$.

- Where p is an AS or PS term and $\mathcal{X}$ an SSV constant, $\llbracket p[\mathcal{X}] \rrbracket^{\langle fmdp,g\rangle} = \llbracket p \rrbracket^{\langle fmdp,g\rangle} \left[\llbracket \mathcal{X} \rrbracket^{\langle fmdp,g\rangle} \right]$ (i.e., the projection of the denotation of p onto the set of variables denoted by $\mathcal{X}$).

- For an AS term s and number n, $\llbracket \mathcal{R}(s) = n \rrbracket^{\langle fmdp,g\rangle} = 1$ iff $\mathcal{R}_{\mathit{fmdp}}(\llbracket s \rrbracket) = n$ (with $\llbracket Q(s,a) > n \rrbracket^{\langle fmdp,g\rangle}$ and $\llbracket Q(s,a) < n \rrbracket^{\langle fmdp,g\rangle}$ defined analogously).

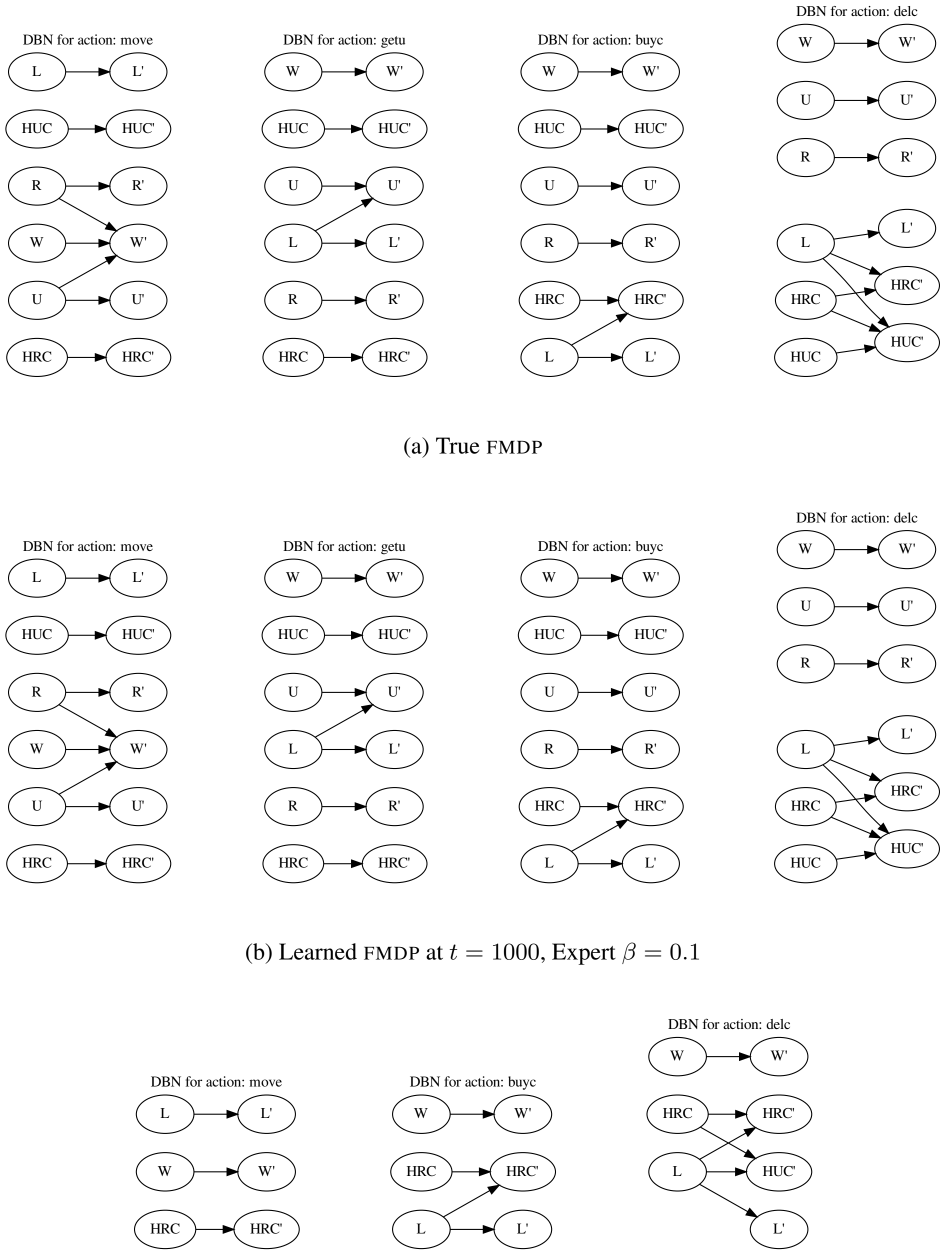
- where p and q are AS, PS, or SV terms, $\llbracket p = q \rrbracket^{\langle fmdp,g\rangle} = 1$ iff $\llbracket p \rrbracket^{\langle fmdp,g\rangle} = \llbracket q \rrbracket^{\langle fmdp,g\rangle}$.

- For formulae ϕ, ψ: $\llbracket \phi \wedge \psi \rrbracket^{\langle fmdp,g\rangle} = 1$ iff $\llbracket \phi \rrbracket^{\langle fmdp,g\rangle} = 1$ and $\llbracket \phi \rrbracket^{\langle fmdp,g\rangle} = 1$; $\llbracket \neg\phi \rrbracket^{\langle fmdp,g\rangle} = 1$ iff $\llbracket \phi \rrbracket^{\langle fmdp,g\rangle} = 0$. Where s is an AS, $\llbracket \exists s \phi \rrbracket^{\langle fmdp,g\rangle} = 1$ iff there is a variable assignment function $g' = g[s/p]$ such that $\llbracket \phi \rrbracket^{\langle fmdp,g'\rangle} = 1$.

- Where V is a SV variable and ϕ is a formula: $\llbracket \exists V \phi \rrbracket^{\langle fmdp,g\rangle} = 1$ iff there exists an SV constant X such that $\llbracket \phi[V/X] \rrbracket^{\langle fmdp,g\rangle} = 1$.

- $\llbracket ?\lambda V \phi \rrbracket^{\langle fmdp,g\rangle} = \{\phi[V/X] \ : \ X \text{ is SV const} \wedge \llbracket \phi[V/X] \rrbracket^{\langle fmdp,g\rangle} = 1\}$.

These interpretations yield a satisfaction relation in the usual way: *fmdp* $\models \phi$ iff there is a function g such that $\llbracket \phi \rrbracket^{\langle fmdp,g\rangle} = 1$.



(a) True FMDP

(b) Learned FMDP at $t = 1000$, Expert $\beta = 0.1$

(c) Learned FMDP at $t = 1000$, Expert $\beta = 0.5$

Figure 1: The true and learned DBN structures on the *coffee-robot* problem.

Agent Type	Better Action	Misunderstanding	Unexpected Reward	Total
Default	75.3	9.8	2.2	87.3
Low Tolerance	168.9	10.9	2.2	182.0
High Tolerance	56.6	8.6	2.3	67.6

Table 1: The average number of messages sent of each type for each setting of expert tolerance in the *factory* problem

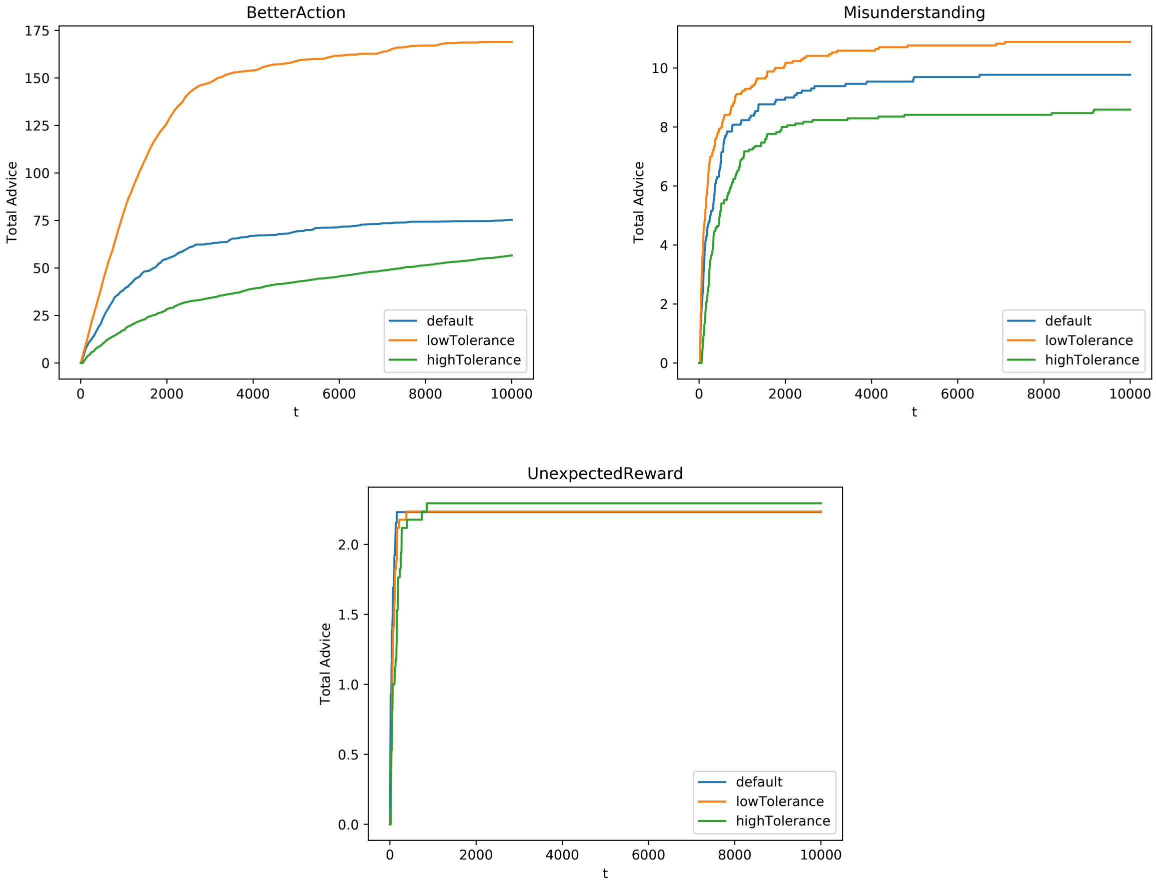


Figure 2: The average number of expert messages sent over time on the *factory* experiment. Results are separated according to advice type