Technical Supplement

1. Proof of Lemmas and Theorems

Lemma 1. *Consider an agent with awareness $A^t \subset A^+$, and expert following (8-13). As $k \to \infty$, either $PE(\pi_{t+k}) \to c \leq \beta$ or the expert utters (13) where $\mathcal{A}' \neq \emptyset$.*

Proof. For finite μ, $(t+k) - t' > \mu$ as $k \to \infty$, satisfying (8).

Given $\mathcal{A}^t$ remains fixed, the agent's policy will eventually converge to some policy π. If $PE(\pi) < \beta$, we are done. If $PE(\pi) \geq \beta$, then (9) will be satisfied.

Since the agent is ϵ-greedy, at every time step it has a non-zero probability of performing all actions $a \in v(\mathcal{A}_t)$, meaning that eventually $\mathcal{A}^e = \mathcal{A}^t$. Further, since $\pi \neq \pi_+$, there must exist some $b \in v(\mathcal{B}^+)$ and $a' \in v(\mathcal{A}^+)$ such that $\forall a \in \mathcal{A}^t, EU(a'|b) > EU(a|b)$, thereby satisfying (11).

Since the agent is ϵ-greedy, it is guaranteed to eventually perform $a^* = \arg\max_{a \in v(\mathcal{A}^t)} EU(a|b)$ in in state b, thus making (12) true for a non-empty value of $\mathcal{A}'$.

(8-12) are not mutually exclusive, so all four will eventually be true at once, causing the expert to utter (13) for non-empty $\mathcal{A}'$. □

Lemma 2. *Consider an agent with awareness $\mathcal{X}^t \subset \mathcal{X}^+$, $\mathcal{A}^t = \mathcal{A}^+$. If $\exists b', \exists b \neq b', b[\mathcal{B}^t] = b'[\mathcal{B}^t]$ and $\pi_+(b) \neq \pi_+(b')$, then as $k \to \infty$, either $PE(\pi_{t+k}) \to c \leq \beta$, or the expert utters (21) such that $B' \notin \mathcal{B}^t$*

Proof. If the agent converges to some policy π such that $PE(\pi) \leq \beta$, we are done.

Assume $a = \pi_+(b)$ and $a' = \pi_+(b')$. Consider that for all k, $P(b_{t+k} = b) > 0$, and (since the agent is ϵ-greedy) $P(\pi_{t+k}(b) = a') > 0$, where $a' = \pi_+(b')$.

If $PE(\pi) \geq$, then at some time $t+k_1$ where $(t+k_1) - t' > \mu$ and $b_{t+k_1} = b$ the agent will perform a', thus satisfying (8-12) and causing the expert to utter $EU(a|w^b_{t+k_1}) > EU(a'|w^b_{t+k_1})$.

By similar reasoning, at some time $b_{t+k_2} = b'$ the expert will utter $EU(a', w^b_{t+k_2}) > EU(a, w^b_{t+k_2})$.

Under $\mathcal{B}^t$, $[\![w^s_{t+k_1}]\!] = [\![w^s_{t+k_2}]\!]$, so the agent will ask (20) with answer $B \notin \mathcal{B}^t$. □

Lemma 3. *Consider an ϵ-greedy agent with awareness $\mathcal{A}^t = \mathcal{A}^+$, $scope_t(\mathcal{R}) \subseteq scope_+(\mathcal{R})$. As $k \to \infty$, there exists a K such that $\forall s, \forall k \geq K$, $\mathcal{R}_{t+k}(s) = \mathcal{R}_+(s)$.*

Proof. Since the $\mathcal{A}^t = \mathcal{A}^+$, and is ϵ-greedy, then over infinite time the agent will eventually enter s at some time i, receiving 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 (22) for s and s' will conflict. The agent resolves this by asking (23), receiving an answer differentiating s from s' in $\mathcal{R}_+$. □

Theorem 1. *Consider an agent with initial awareness $\mathcal{X}^0 \subseteq \mathcal{X}^+, A^0 \subseteq A^+, scope_0(\mathcal{R}) \subseteq scope_+(\mathcal{R})$ following algorithm 1 (with $\kappa = 0$). As $t \to \infty$, $PE(\pi_t) \to c \leq \beta$.*

Proof. By repeatedly applying theorems 1-3, either $PE(\pi_t) \to c \leq$ (in which case we are done), or there exists a K where $\mathcal{A}^K = \mathcal{A}^+$, $\mathcal{R}_K = \mathcal{R}_+$, and $\mathcal{X}^K = \mathcal{B}^k \cup \mathcal{O}^K$. Here, $\mathcal{B}^K$ contains all variables $\mathcal{B} \in \mathcal{B}^+$ such that there exist b, and b' where $b[\mathcal{B}^+ \setminus B] = b'[\mathcal{B}^+ \setminus B]$ but $b[B] \neq b'[B]$ and $\pi_+(b) \neq \pi_+(b')$. In other words, $\langle \mathcal{X}^K, \mathcal{A}^K, \mathcal{R}^K \rangle$ define a related decision problem with an identical optimal policy and marginal probability distribution to the original problem, but for which the agent is fully aware.

The BD-score is a *consistent score*, which means that as $|D| \to \infty$, Pa^* contains all dependencies present in the true distribution, regardless of initial prior (assuming we have not pruned the search space, which is true in this case because $\kappa = 0$). As a result, as $t \to \infty$, then $\forall a \in v(\mathcal{A}^+), \forall b \in \mathcal{B}^K$, $\forall y \in v(scope_+(\mathcal{R}))$, we have that $P(y|a, b, Pa^*_t, \theta^*_t) \to P(y|a, b, Pa^+, \theta^+)$.

Since our agents probabilistic model converges to the true probabilistic distribution, and because $\mathcal{R}^K = \mathcal{R}^+$ and $\mathcal{A}^K = \mathcal{A}^+$, we have that $\pi^*_t \to \pi_+$ and therefore that $PE(\pi^*_t) \to 0$ □

2. Derivation of Equation (7)

Note, the derivation below corrects an error from the original Buntine (1991) paper.

The proof that:

$$\mathrm{BD}_t(X, Pa_X) = \mathrm{BD}_{t-1}(X, Pa_X)\frac{N^t_{i|j} + \alpha_{i|j} - 1}{N^t_{.|j} + \alpha_{.|j} - 1} \tag{7}$$

Follows from the definition of the multivariate-β function in terms of the Γ-function, and the recursive structure of Γ:

$$\beta(n_1, \dots n_m) = \frac{\prod_{k=0}^{m} \Gamma(n_k)}{\Gamma(\sum_{k=0}^{m} n_k)} \tag{1}$$

$$\Gamma(x) = (x-1)\Gamma(x-1) \tag{2}$$

From equation (3), we have:

$$\begin{aligned}&\mathrm{BD}_t(X, Pa_X) = \\ &\quad P(Pa_X)\prod_{j \in v(Pa_X)} \frac{\beta(N^t_{0|j} + \alpha_{0|j}, \dots, N^t_{m|j} + \alpha_{m|j})}{\beta(\alpha_{0|j}, \dots, \alpha_{m|j})}\end{aligned} \tag{3}$$

Lets assume that $d_t[X] = i$ and $d_t[Pa_X] = j$. The only difference between $BD_t(X, Pa_X)$ and $BD_{t-1}(X, Pa_X)$ is between counts $N^t_{i|j}$ and $N^{t-1}_{i|j}$ (specifically, that $N^{t-1}_{i|j} = N^t_{i|j} - 1$), so we can rewrite (3) as:

$$\begin{aligned}&\mathrm{BD}_t(X, Pa_X) \\ &= BD_{t-1}(X, Pa)\frac{\Gamma(\sum_{k=0}^{m} N^{t-1}_{k|j})\prod_{k=0}^{m}\Gamma(N^t_{k|j})}{\Gamma(\sum_{k=0}^{m} N^t_{k|j})\prod_{k=0}^{m}\Gamma(N^{t-1}_{k|j})} \\ &= \frac{\Gamma((\sum_{k=0}^{m} N^t_{k|j}) - 1)\Gamma(N^t_{i|j})}{\Gamma(\sum_{k=0}^{m} N^t_{k|j})\Gamma(N^t_{i|j} - 1)}\end{aligned}$$

Then, via (2), we have:

$$\begin{aligned}&\frac{\Gamma((\sum_{k=0}^{m} N^t_{k|j}) - 1)\Gamma(N^t_{i|j})}{\Gamma(\sum_{k=0}^{m} N^t_{k|j})\Gamma(N^t_{i|j} - 1)} \\ &= \frac{\Gamma((\sum_{k=0}^{m} N^t_{k|j}) - 1)(N^t_{i|j} - 1)\Gamma(N^t_{i|j} - 1)}{((\sum_{k=0}^{m} N^t_{k|j}) - 1)\Gamma((\sum_{k=0}^{m} N^t_{k|j}) - 1)\Gamma(N^t_{i|j} - 1)} \\ &= \frac{N^t_{i|j} - 1}{(\sum_{k=0}^{m} N^t_{k|j}) - 1}\end{aligned}$$

And therefore that:

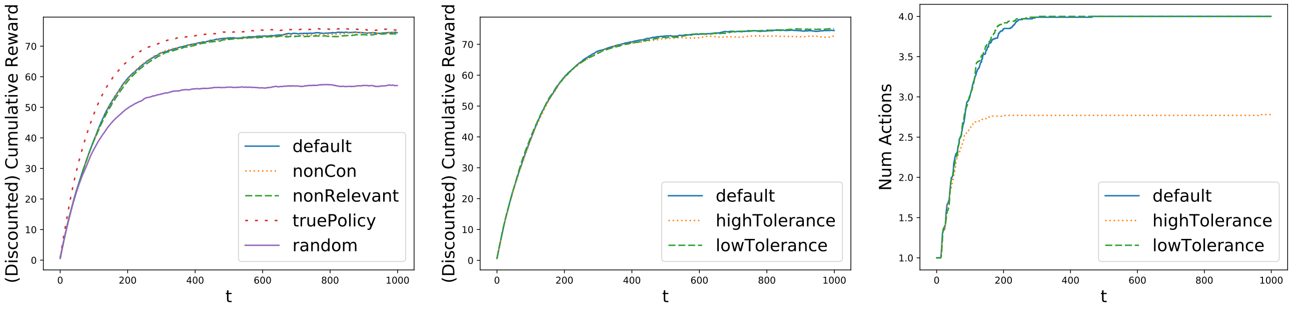
$$\mathrm{BD}_t(X, Pa_X) = \mathrm{BD}_{t-1}(X, Pa_X)\frac{N^t_{i|j} + \alpha_{i|j} - 1}{N^t_{.|j} + \alpha_{.|j} - 1} \tag{7}$$

As required.

3. Specifications of IDs from Experiments + Additional Results

Tables 1-3 below provide a full specification of the three randomly generated IDs used in the Experiments Section of the main paper. All variables are binary. The numbers in the $P(X|Pa)$ give the probability of $X = 1$ for each possible assignment $v(Pa)$, where the first probability refers to the assignment of all parents to 0, and the last probability refers to the assignment of all parents to 1.

Figures 1 and 2 give the results give the results of our experiments on the small (12 variable) and medium (24 variable) IDs, using the same experimental setup as presented in Section 4 of the main paper. As with the main paper, the default agent is able to successfully converge on the true optimal policy, despite starting out unaware of factors critical to success. However, the differences in varying the expert tolerances and between the conservative and non-conservative agents are less pronounced, simply because the underlying decision problem is simpler in the smaller cases.

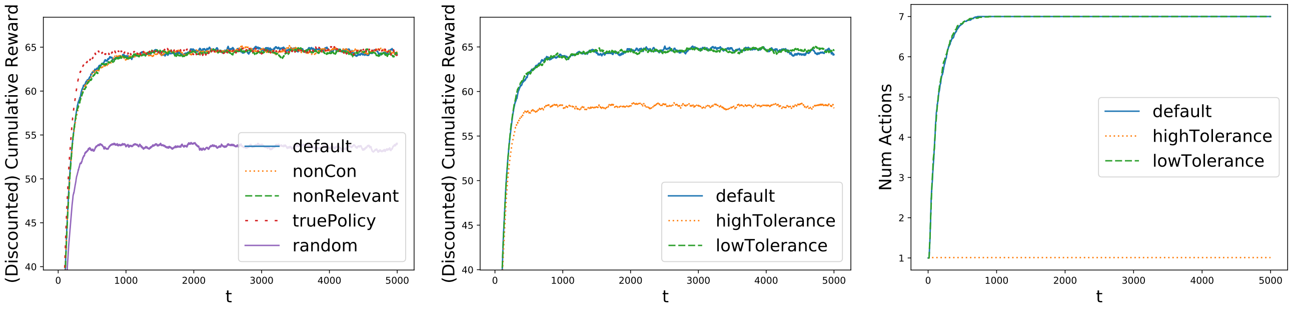


(a)

(b)

(c)

Figure 1. Rewards and size of $|\mathcal{A}|$ over time on id-small.



(a)

(b)

(c)

Figure 2. Rewards and size of $|\mathcal{A}|$ over time on id-medium.

X	Pa_X	$P(X\|Pa_X)$
B1	$\emptyset$	0.50
O1	A4	0.23, 0.18
B2	$\emptyset$	0.84
B3	$\emptyset$	0.18
O3	A1, A3, B3	0.31, 0.89, 0.87, 0.15, 0.48, 0.21, 0.90, 0.06
B4	B2	0.45, 0.26
O4	A2, B1	0.78, 0.87, 0.12, 0.07
O2	B2, B4, O1	0.04, 0.75, 0.64, 0.07, 0.36, 0.62, 0.69, 0.17

	$scope(\mathcal{R})$	$\mathcal{R}(s)$
	O4, O2, O3	0.98, 0.14, 0.70, 0.35, 0.38, 0.98, 0.11, 0.97

Table 1. id-small: (12 Variables)

X	Pa_X	$P(X\|Pa_X)$
B8	$\emptyset$	0.05
B6	$\emptyset$	0.97
B7	$\emptyset$	0.01
B1	$\emptyset$	0.72
B4	$\emptyset$	0.08
B5	$\emptyset$	0.66
B2	$\emptyset$	0.78
B3	$\emptyset$	0.54
O3	A6, B4, B6	0.66, 0.14, 0.99, 0.83, 0.75, 0.71, 0.30, 0.12
O2	A5, A8, B5	0.26, 0.94, 0.67, 0.18, 0.04, 0.63, 0.87, 0.14
O6	A6, B3	0.87, 0.31, 0.48, 0.13
O4	A1, B8	0.27, 0.51, 0.12, 0.73
O8	A4, B2, O4	0.81, 0.51, 0.44, 0.44, 0.34, 0.82, 0.04, 0.56
O1	A2, A7, O3	0.03, 0.47, 0.74, 0.43, 0.59, 0.08, 0.60, 0.57
O7	B7, O6	0.83, 0.68, 0.27, 0.23
O5	A3, B1, O7	0.75, 0.30, 0.25, 0.35, 0.64, 0.30, 0.31, 0.31

	$scope(\mathcal{R})$	$\mathcal{R}$
	O1, O8, O5, O2	0.47, 0.95, 1.00, 0.18, 0.26, 0.75, 0.38, 0.76, 0.60, 0.78, 0.02, 0.75, 0.40, 0.32, 0.30, 0.14

Table 2. id-medium (24 variables)

X	Pa_X	$P(X\|Pa_X)$
B10	$\emptyset$	0.43
O9	A12	0.53, 0.82
B11	$\emptyset$	0.59
O8	A2	0.87, 0.19
B12	$\emptyset$	0.42
B13	$\emptyset$	0.52
B14	$\emptyset$	0.76
B15	$\emptyset$	0.12
O2	A1, A6, A8	0.43, 0.89, 0.00, 0.18, 0.19, 0.76, 0.76, 0.67
B8	$\emptyset$	0.28
B9	$\emptyset$	0.26
B6	$\emptyset$	0.96
B7	$\emptyset$	0.27
B1	$\emptyset$	0.99
B4	$\emptyset$	0.99
B5	$\emptyset$	0.13
B2	$\emptyset$	0.77
B3	$\emptyset$	0.01
O3	B2, B9, O2	0.31, 0.21, 0.71, 0.65, 0.78, 0.17, 0.67, 0.92
O6	A3, B6, B11	0.28, 0.41, 0.55, 0.42, 0.93, 0.58, 0.44, 0.74
O5	A13, B3, B4	0.17, 0.66, 0.93, 0.62, 0.47, 0.08, 0.82, 0.86
O15	A2, A11, O8	0.92, 0.34, 0.25, 0.48, 0.43, 0.25, 0.52, 0.32
O14	A10, B5, B8	0.37, 0.91, 0.75, 0.86, 0.29, 0.33, 0.96, 0.84
O13	A4, A7, B1	0.37, 0.40, 0.74, 0.94, 0.67, 0.70, 0.66, 0.46
O7	B7, B15, O6	0.31, 0.69, 0.17, 0.56, 0.88, 0.40, 0.65, 0.91
O4	A14, O3, O9	0.31, 1.00, 0.59, 0.81, 0.08, 0.20, 0.02, 0.63
O12	B13, B14, O13	0.55, 0.71, 0.61, 0.22, 0.51, 0.90, 0.37, 0.52
O11	A5, B10, O5	0.73, 0.65, 0.42, 0.47, 0.42, 0.34, 0.97, 0.33
O10	A15, B12, O15	0.06, 0.80, 0.42, 0.36, 0.62, 0.97, 0.15, 0.29
O1	A9, O7, O11	0.48, 0.03, 0.14, 0.41, 0.16, 0.64, 0.22, 0.19

	$scope(\mathcal{R})$	$\mathcal{R}$
	O1, O4, O14, O10, O12	0.28, 0.19, 0.91, 0.71, 0.68, 0.57, 0.98, 0.86, 0.96, 0.54, 0.19, 0.20, 0.72, 0.61, 0.32, 0.65, 0.64, 0.12, 0.64, 0.67, 0.72, 0.27, 0.05, 0.07, 0.63, 0.09, 0.55, 0.45, 0.19, 0.02, 0.51, 0.13

Table 3. id-large (36 variables)