Algorithm 1 HumanPolicyGenerator

Input: PolicyFunction

Output: $\pi^{(2)}_{\textsc{PolicyFunction}}$

1: Start from $\pi^{(2)}_{\textsc{PolicyFunction}}(s) = \textsc{NaN}, \forall s \in S$
2: **for** $s \in S$ **do**
3: $\quad \pi^{(2)}_{\textsc{PolicyFunction}}(s) = \textsc{PolicyFunction}(s)$
4: **end for**

Algorithm PolicyFunction:Table

Input: $s \in S$

Output: $a \in A$

Algorithm PolicyFunction:Below

Input: $s \in S, \{\textsc{left}, \textsc{right}\}, \{\textsc{left}, \textsc{right}\}$

Output: $a \in A$

Algorithm 2 ValueIteration

Input: $\pi^{(2)}$

Output: $\pi^{(1)}$, $V(\pi^{(1)}, \pi^{(2)})$

Algorithm 3 Q learning with UCB exploration

Input: ε, ε_2, δ, c_2, and ITERS.
Output: $Q\colon S \times A \to [0,1]$.

$Q_0(s,a), \hat{Q}_0(s,a) \leftarrow \frac{1}{1-\gamma}$, $N(s,a)=0$, $R = \lceil log(\frac{1}{\varepsilon_2(1-\gamma)})/(1-\gamma)\rceil$.
$L = \lfloor log_2 R \rfloor$, $\varepsilon_L = \frac{1}{2^{L+2}}\varepsilon_2(log(1/(1-\gamma)))^{-1}$.
$M = \max\{\lceil 2log_2(\frac{1}{\varepsilon_L(1-\gamma)})\rceil, 10\}$, $\varepsilon_1 \leftarrow \frac{\varepsilon}{24RMlog\frac{1}{1-\gamma}}$, $H = \frac{log1/((1-\gamma)\varepsilon_1)}{log1/\gamma}$.
$\iota(k) \triangleq log(SA(k+1)(k+2)/\delta)$, $\alpha_k \triangleq \frac{H+1}{H+k}$.
for $i = 1\colon$ ITERS **do**
    $a_i \leftarrow \arg\max_{a'} \hat{Q}(s_i, a')$.
    Receive reward r_i and transit to s_{i+1}.
    $N(s_i, a_i) \leftarrow N(s_i, a_i) + 1$.
    $k \leftarrow N(s_i, a_i)$, $b_k = \frac{c_2}{1-\gamma}\sqrt{\frac{H\iota(k)}{k}}$
    $Q_{i+1}(s,a) \leftarrow (1-\alpha)Q_i(s,a) + \alpha[r(s,a) + \gamma \max_{a'} \hat{Q}_{i+1}(s',a') + b_k]$.
    $\hat{Q}_{i+1}(s,a) \leftarrow \min[\hat{Q}_{i+1}(s,a), Q_i(s,a)]$.
end for

Algorithm 4 Maximize to Explore

Input: η, Hypothesis class $\mathcal{H}$. VALUEITERATION.
Output: $\pi^{(1)}$

$\mathcal{D} = \phi$
for $i = 1\colon$ ITERS **do**
    $\hat{\pi} = \arg\max_{\pi^{(2)} \in \mathcal{H}}(V(\text{VALUEITERATION}(\pi^{(2)}), \pi^{(2)}) - \eta\mathcal{L}^{\pi^{(2)}}(\mathcal{D}))$
    $\pi^{(1)} \leftarrow \text{VALUEITERATION}(\hat{\pi})$
    Simulate with $\pi^{(1)}$ until the game terminates and store (s,a,s') pairs into $\mathcal{D}$.
end for

Algorithm 5 Upper Confidence Bound

Input: α, Hypothesis class $\mathcal{H}$, VALUEITERATION.
Output: $\pi^{(1)}$.

$N(\pi) \leftarrow 1$ and $V(\pi) \leftarrow 0$ for all π.
for $i = 1\colon$ ITERS **do**
    $\hat{\pi} = \arg\max_{\pi^{(1)} \in \text{VALUEITERATION}(\mathcal{H})}(V(\pi^{(1)}) + \sqrt{\frac{\alpha logi}{N(\pi^{(1)})}})$.
    $\pi^{(1)} \leftarrow \hat{\pi}$
    $N(\pi^{(1)}) \leftarrow N(\pi^{(1)}) + 1$.
    Simulate with $\pi^{(1)}$ and receive the cumulative reward V.
    $V(\pi^{(1)}) \leftarrow V$.
end for