**Algorithm 1** Preference-Based Learning of Reward Functions

```
Input: Features $\phi$, horizon $N$, dynamics $f$, $iter$
Output: Distribution of $\mathbf{w}$: $p(\mathbf{w})$
Initialize $p(\mathbf{w}) \sim \text{Uniform}(B)$, for a unit ball $B$
While $t < iter$:
    $W \leftarrow M$ samples from $\text{AdaptiveMetropolis}(p(\mathbf{w}))$
    $(x^0, \mathbf{u}_R, \mathbf{u}_H^A, \mathbf{u}_H^B) \leftarrow \text{SynthExps}(W, f)$
    $I_t \leftarrow \text{QueryHuman}(x^0, \mathbf{u}_R, \mathbf{u}_H^A, \mathbf{u}_H^B)$
    $\varphi = \Phi(x^0, \mathbf{u}_R, \mathbf{u}_H^A) - \Phi(x^0, \mathbf{u}_R, \mathbf{u}_H^B)$
    $f_\varphi(\mathbf{w}) = \min(1, I_t \exp(\mathbf{w}^\top \varphi))$
    $p(\mathbf{w}) \leftarrow p(\mathbf{w}) \cdot f_\varphi(\mathbf{w})$
    $t \leftarrow t + 1$
End for
```