CS 690: Human-Centric Machine Learning

Prof. Scott Niekum

Interactive RL

Reward design is hard, demonstrations are hard

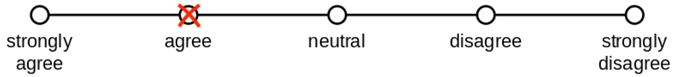
What about feedback?

Feedback: Clicker training

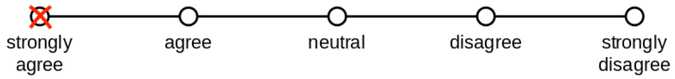
Feedback: Likert ratings

Website User Survey

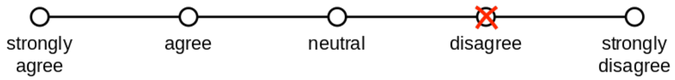
1. The website has a user friendly interface.



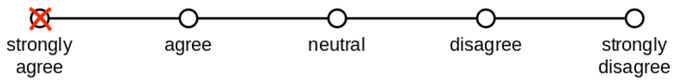
2. The website is easy to navigate.



3. The website's pages generally have good images.



4. The website allows users to upload pictures easily.

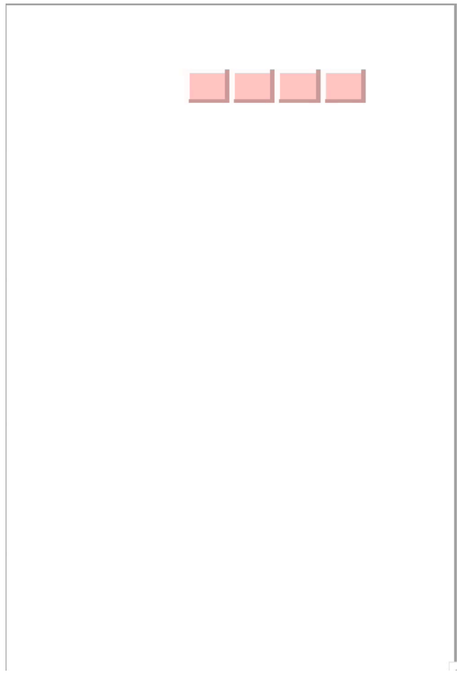


5. The website has a pleasing color scheme.

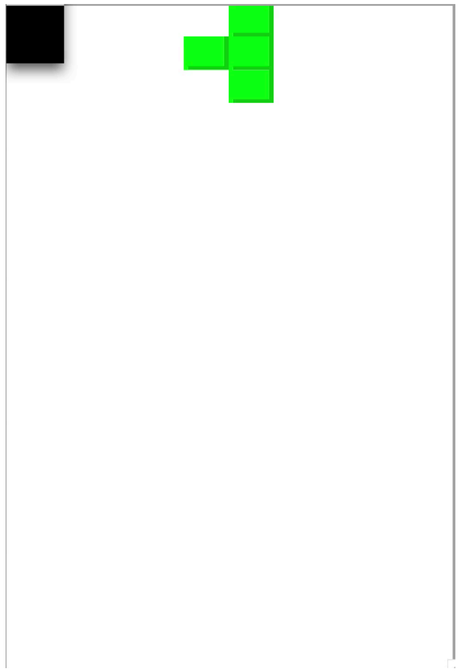
Feedback: Implicit feedback

TAMER

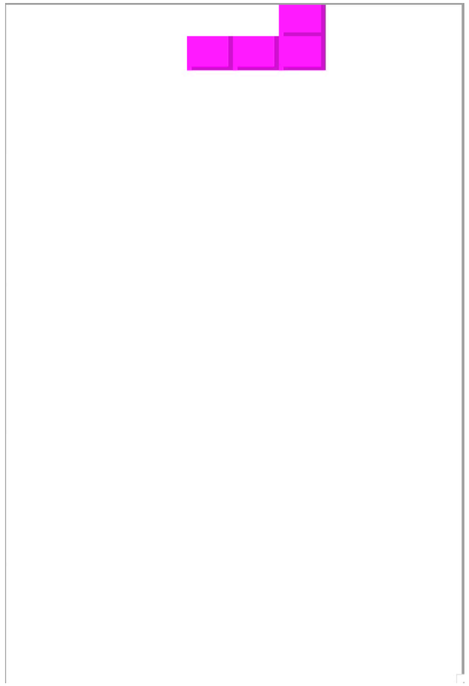
Before training:



Training:

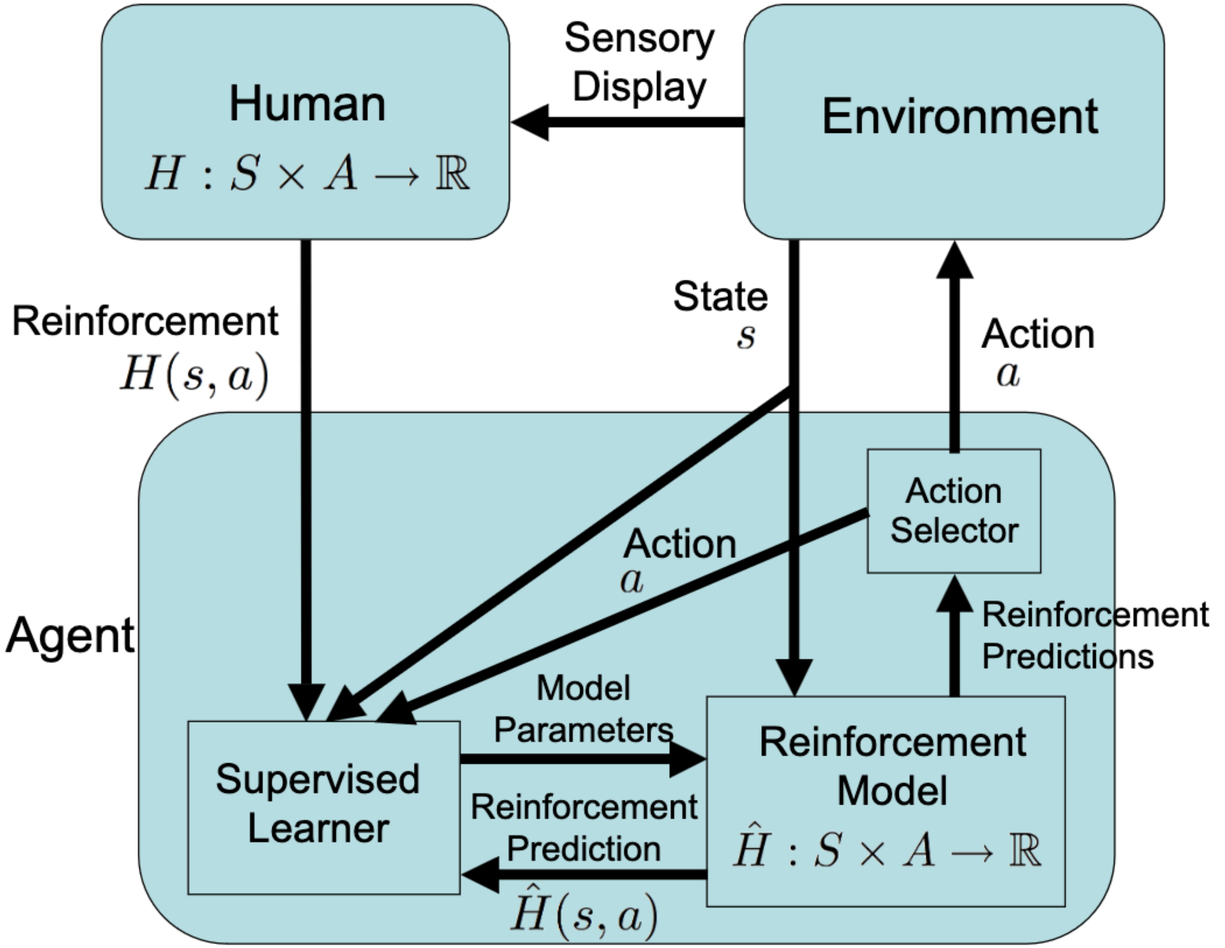


After training:



Knox and Stone, ICDL 2008

TAMER


Human
$H : S \times A \rightarrow \mathbb{R}$
Sensory
Display
Environment
Reinforcement
$H(s, a)$
State
s
Action
a
Agent
Action
a
Action
Selector
Reinforcement
Predictions
Model
Parameters
Supervised
Learner
Reinforcement
Model
$\hat{H} : S \times A \rightarrow \mathbb{R}$
Reinforcement
Prediction
$\hat{H}(s, a)$

TAMER

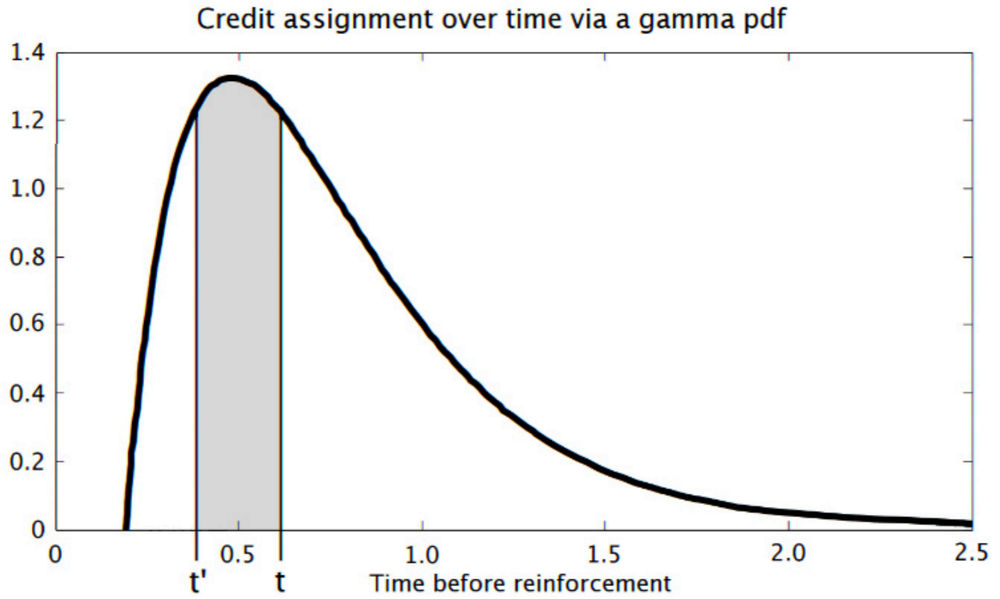
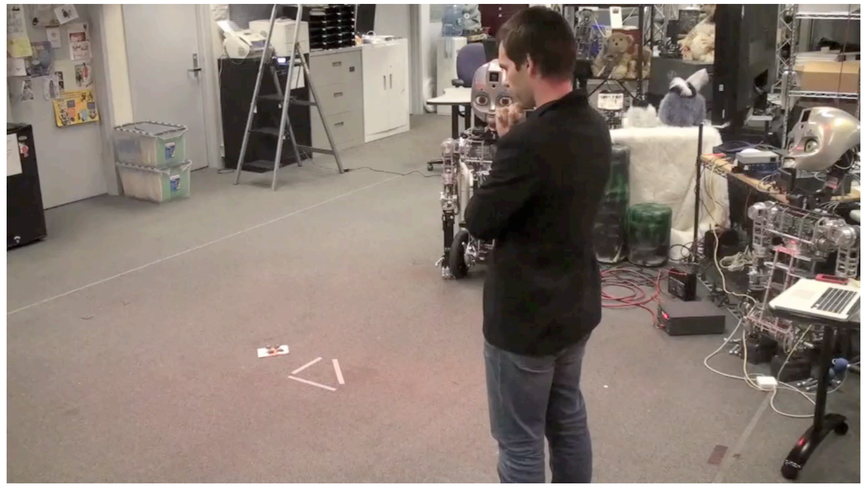


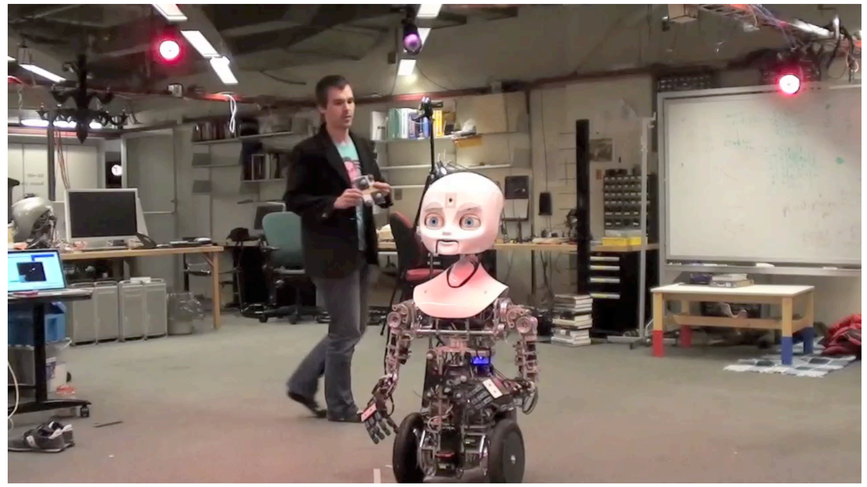
Figure 3: Probability density function $f(x)$ for a gamma(2.0, 0.28) distribution. Reinforcement signal h is received at time 0. If t and t' are times of consecutive time steps, credit for the time step at t is $\int_{t'}^{t} f(x)dx$. Note that time moves backwards as one moves right along the x-axis.

TAMER

Go to

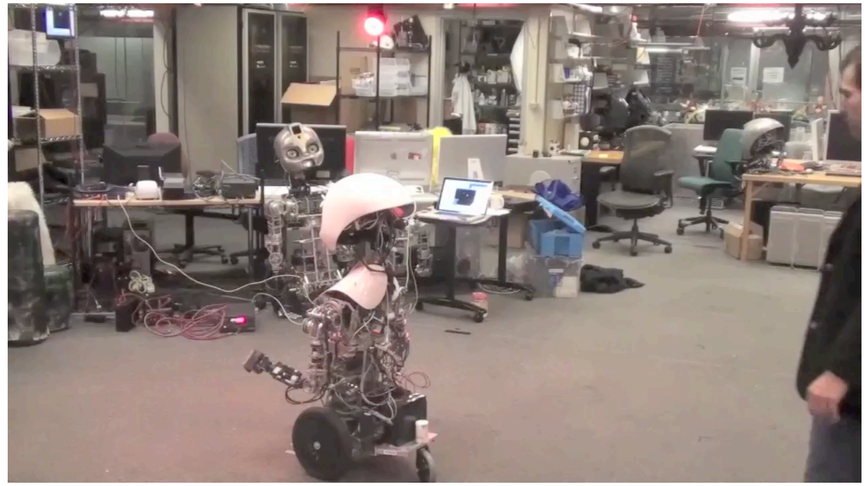


Look away

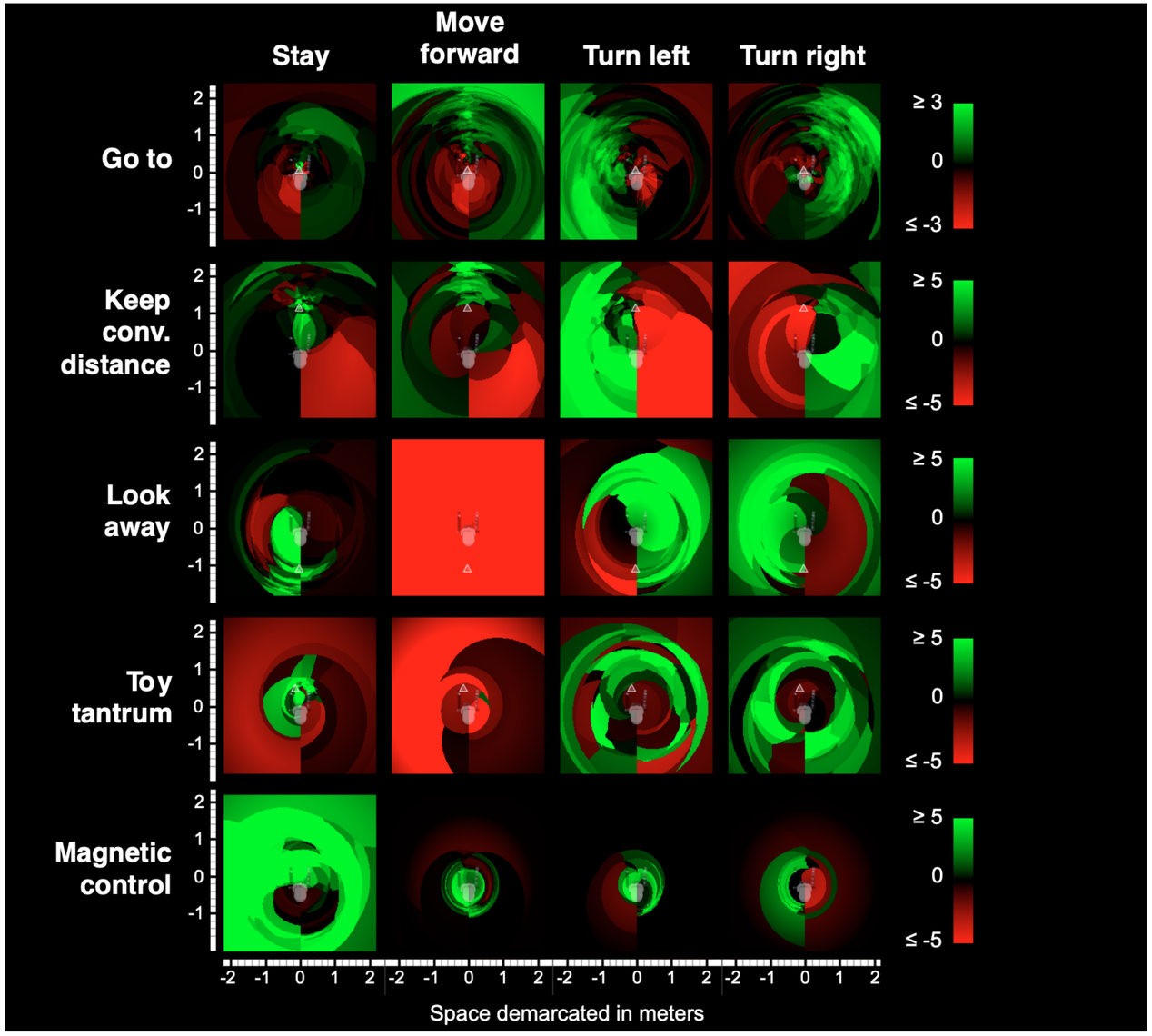


Toy tantrum

Magnetic control



TAMER


Stay
Move forward
Turn left
Turn right
Go to
Keep conv. distance
Look away
Toy tantrum
Magnetic control
2
1
0
-1
≥ 3
0
≤ -3
≥ 5
0
≤ -5
-2 -1 0 1 2
Space demarcated in meters

COACH

TAMER feedback interpretation: (correlated with) $Q^*(s, a)$

COACH feedback interpretation: (correlated with) $A^\pi(s, a)$

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s).$$

MacGlashan, James, et al. "Interactive learning from policy-dependent human feedback." *International conference on machine learning*. PMLR, 2017.

COACH

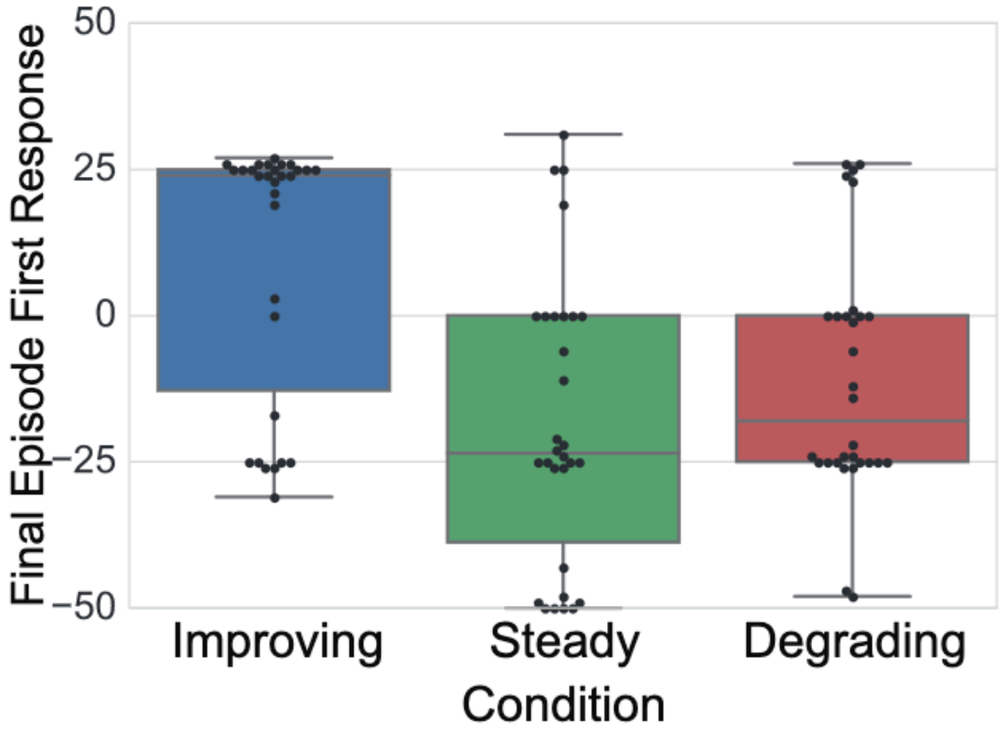


Figure 2. The feedback distribution for first step of the final episode for each condition. Feedback tended to be positive for improving behavior, but negative otherwise.

COACH

$$\Delta\theta = \alpha \nabla_\theta \rho$$

$$\begin{aligned}
\nabla_\theta \rho &= \sum_s d^\pi(s) \sum_a \nabla_\theta \pi(s,a) Q^\pi(s,a) \\
&= \sum_s d^\pi(s) \sum_a \pi(s,a) \nabla_\theta \pi(s,a) \frac{Q^\pi(s,a)}{\pi(s,a)} \\
&= E_\pi \left[\nabla_\theta \pi(s,a) \frac{Q^\pi(s,a)}{\pi(s,a)} \right]
\end{aligned}$$

$$\boldsymbol{\Delta\theta_t} = \alpha_t \nabla_{\boldsymbol{\theta}} \pi(s_t, a_t) \frac{f_{t+1}}{\pi(s_t, a_t)}$$

COACH

Algorithm 1 Real-time COACH

Require: policy $\pi_{\boldsymbol{\theta}_0}$, trace set $\boldsymbol{\lambda}$, delay d, learning rate α

Initialize traces $\boldsymbol{e}_\lambda \leftarrow \mathbf{0}\ \forall \lambda \in \boldsymbol{\lambda}$

observe initial state s_0

for $t = 0$ to ∞ **do**

select and execute action $a_t \sim \pi_{\boldsymbol{\theta}_t}(s_t, \cdot)$

observe next state s_{t+1}, sum feedback f_{t+1}, and λ

for $\lambda' \in \boldsymbol{\lambda}$ **do**

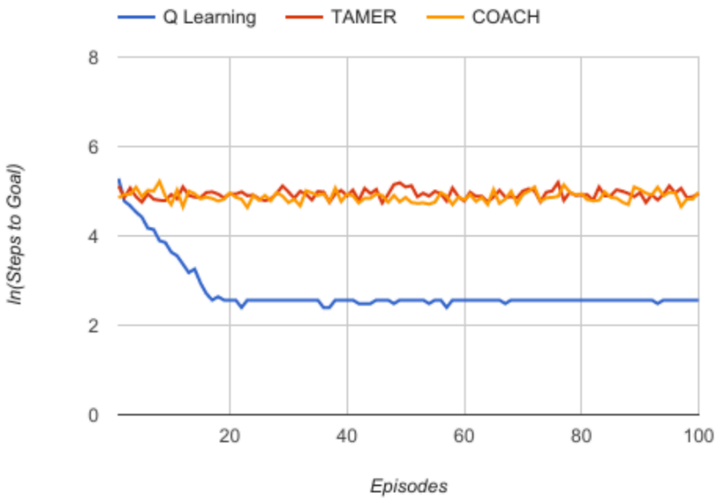
$$\boldsymbol{e}_{\lambda'} \leftarrow \lambda' \boldsymbol{e}_{\lambda'} + \frac{1}{\pi_{\boldsymbol{\theta}_t}(s_{t-d}, a_{t-d})} \nabla_{\boldsymbol{\theta}_t} \pi_{\boldsymbol{\theta}_t}(s_{t-d}, a_{t-d})$$

end for

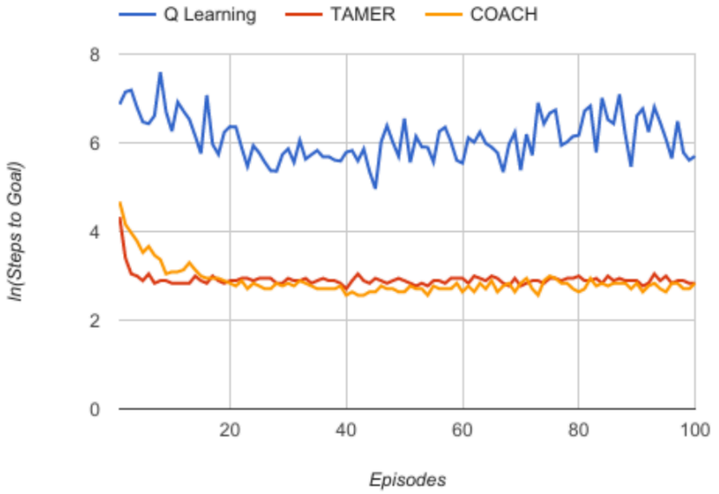
$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \alpha f_{t+1} \boldsymbol{e}_\lambda$$

end for

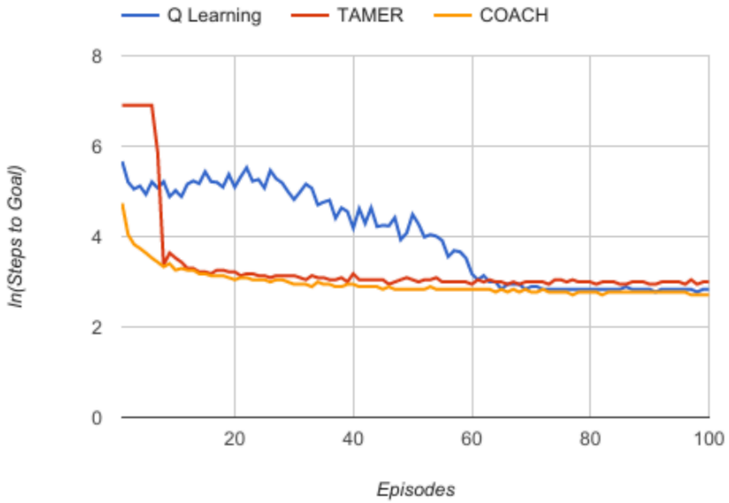
COACH



(a) Task feedback



(b) Action feedback



(c) Improvement feedback

COACH