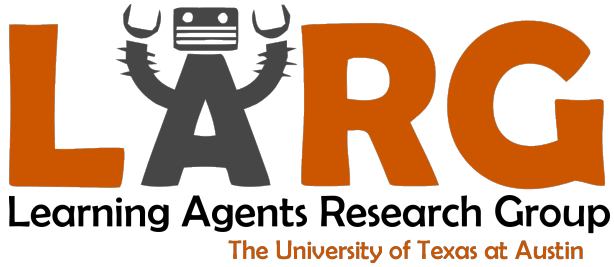
Understanding Human Teaching Modalities in Reinforcement Learning Environments

A Preliminary Report

Slides available on the Program page of the ALIHT website.

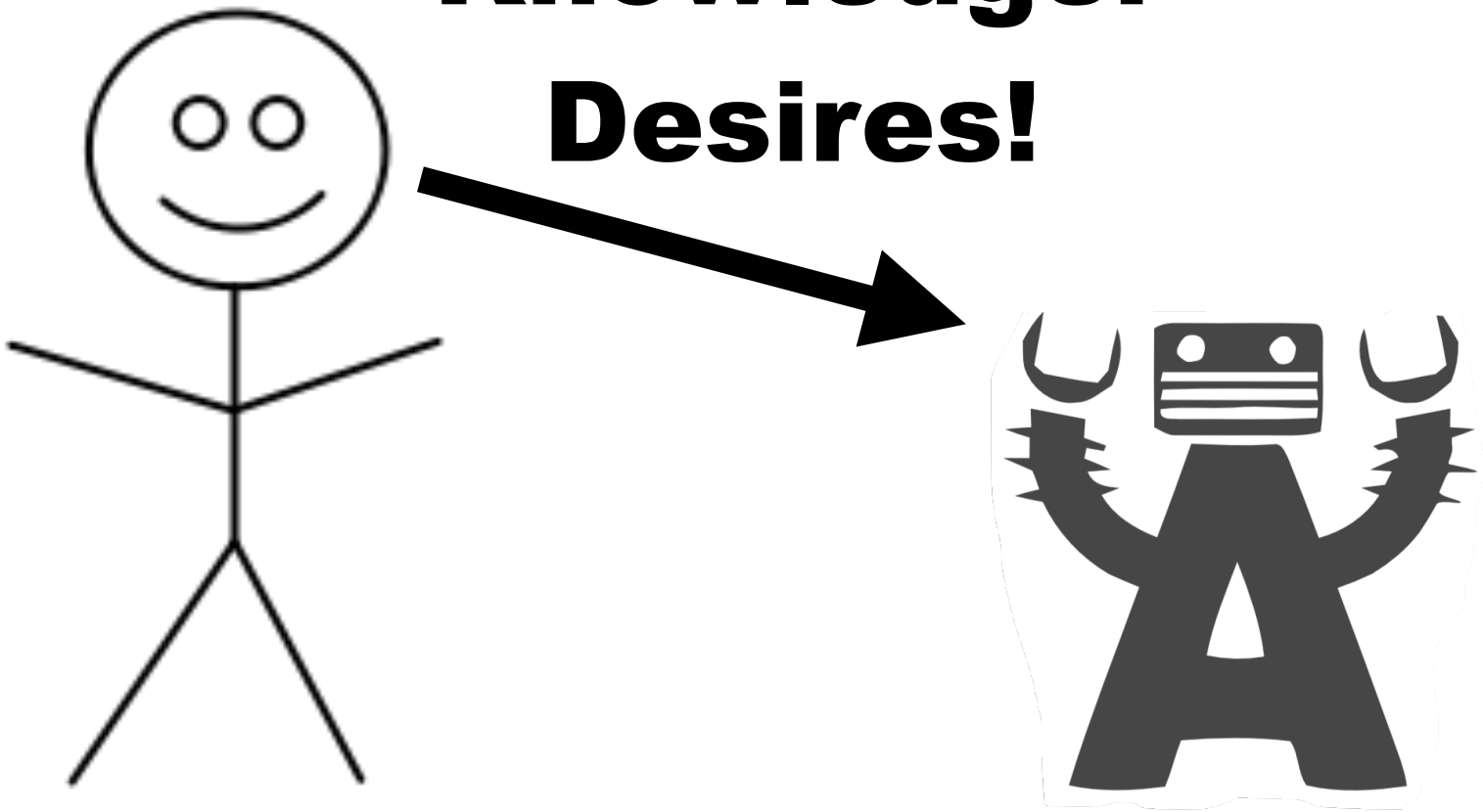
W. Bradley Knox
and **Peter Stone**



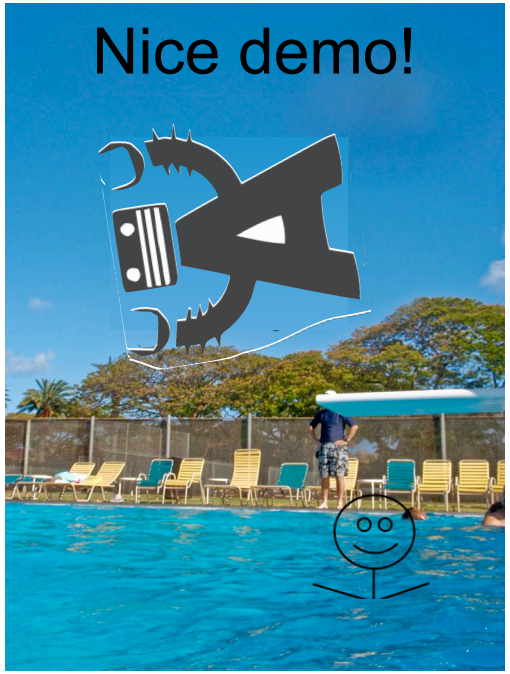
Matthew E. Taylor

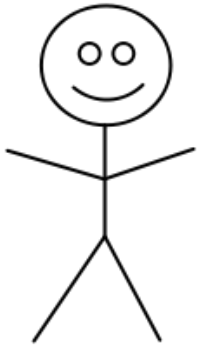
Knowledge!

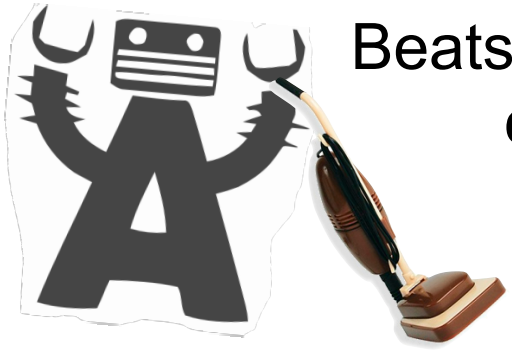
Desires!



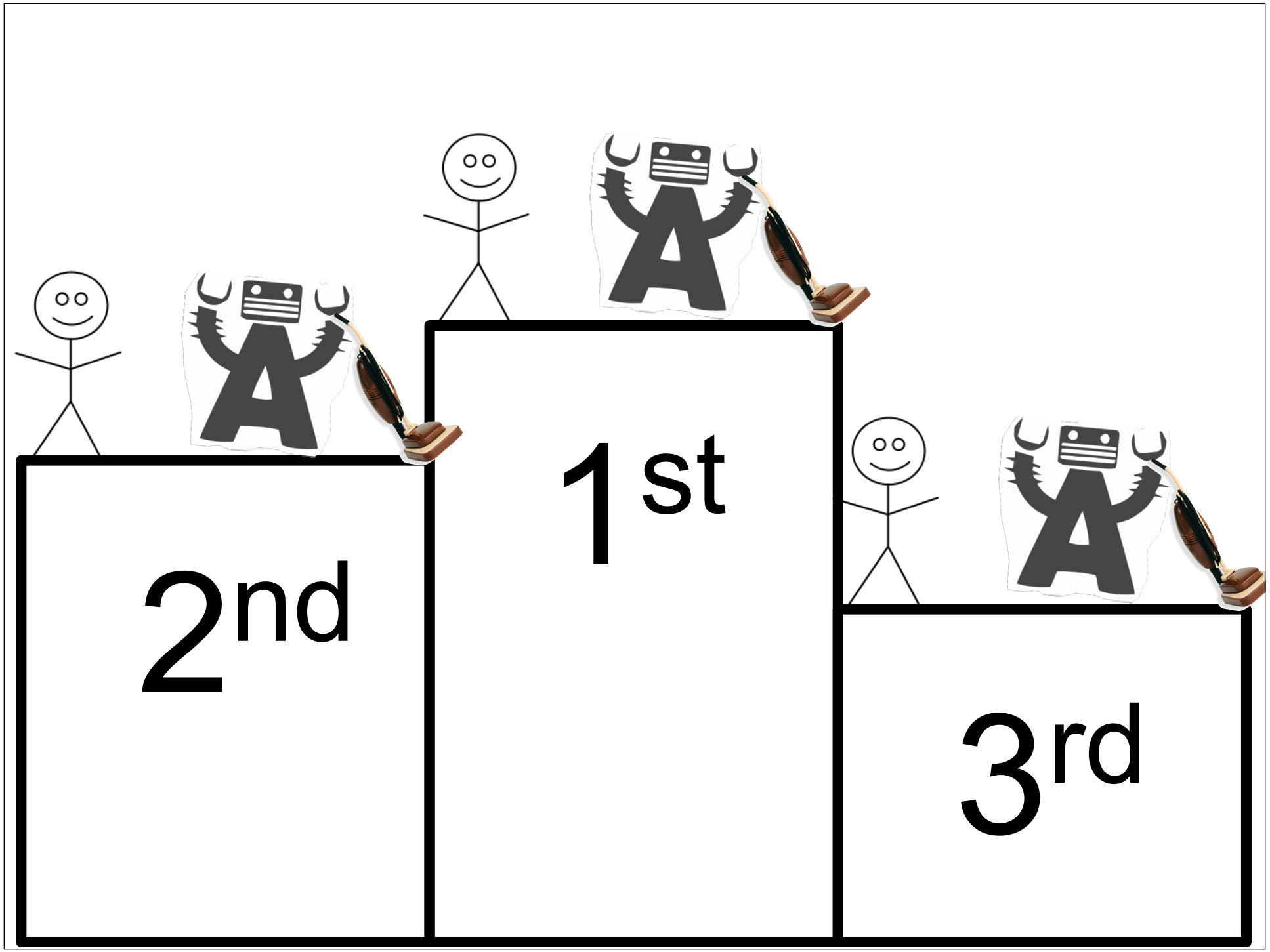
Current state of interactive learning evaluation





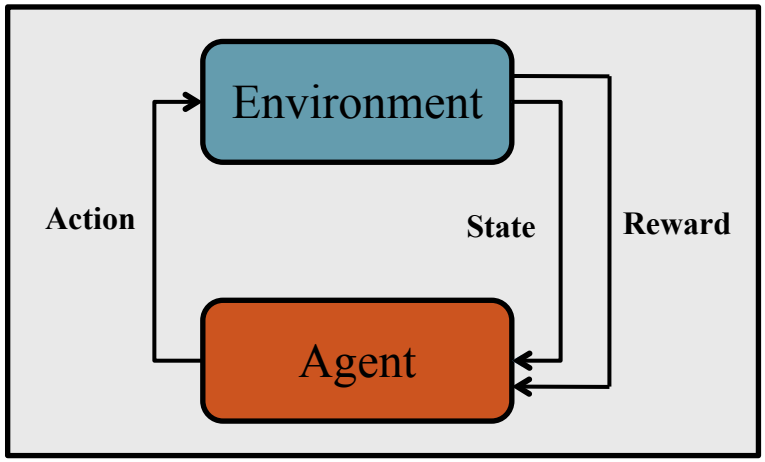


Beats hand-coded!


2nd
1st
3rd

Reinforcement learning tasks

- Learn from limited feedback
- Delayed reward
- Very general
  - Possibly slow learning
- Human end-user cannot determine correct behavior



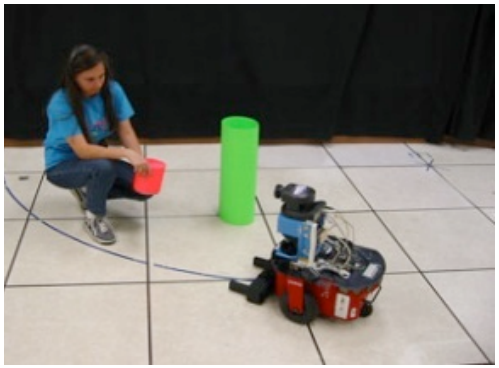
Learning from demonstration (LfD)

- Goal: reproduce behavior / policy
  - generalizing effectively to unseen situations
- Argall, Chernova, Veloso and Browning. **A Survey of Robot Learning from Demonstration.** *RAS, 2009.*

Lockerd & Breazeal

Grollman & Jenkins

Argall, Browning & Veloso



Nicolescu & Matarić

Learning from feedback (interactive shaping)

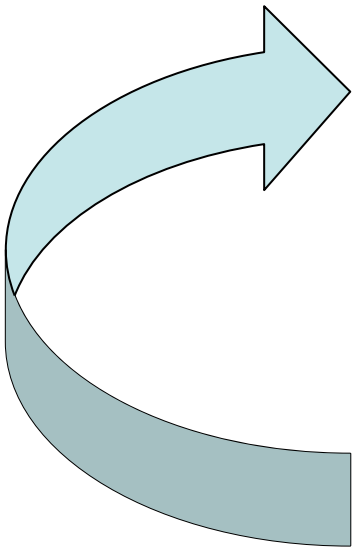
TAMER

Key insight: trainer evaluates behavior using a model of its long-term quality

Knox and Stone, K-CAP 2009

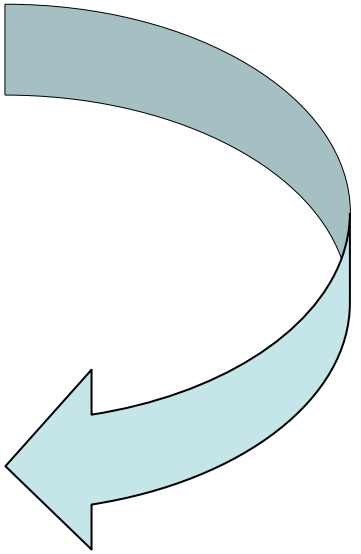
Learning from feedback (interactive shaping)

TAMER



Learn a model of human reinforcement

$$H : S \times A \rightarrow \mathbb{R}$$



Directly exploit the model to determine action

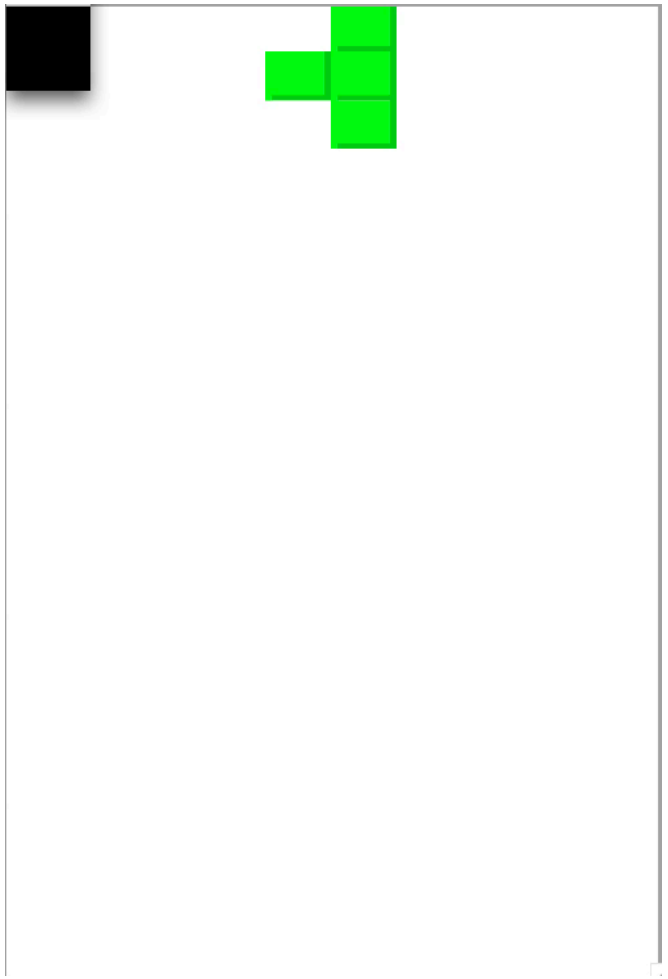
If greedy:

$$action = argmax_a \hat{H}(s, a)$$

Knox and Stone, K-CAP 2009

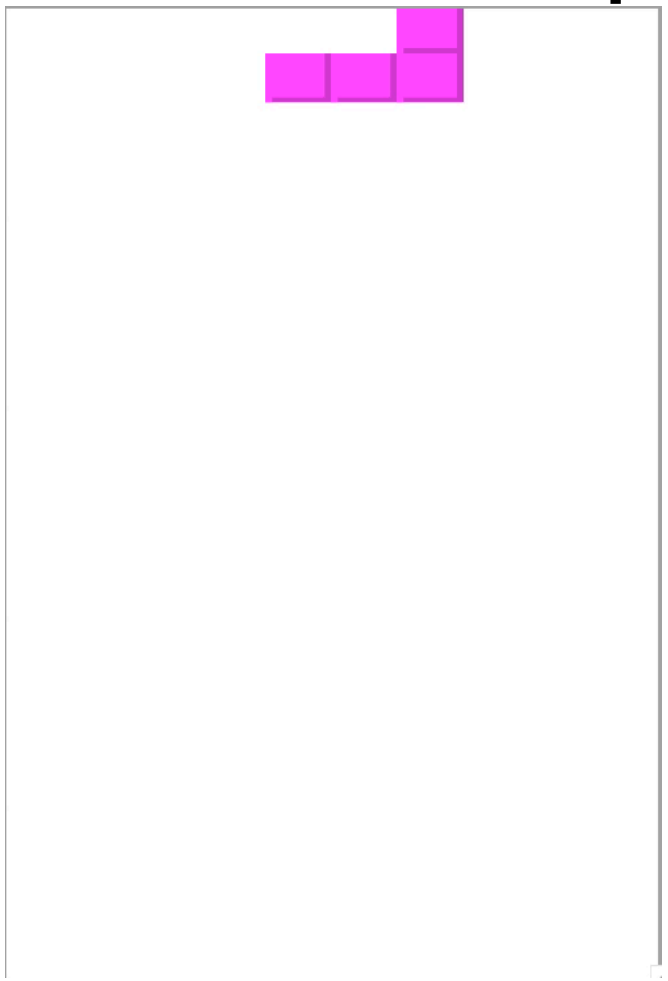
Learning from feedback (interactive shaping)

Training:



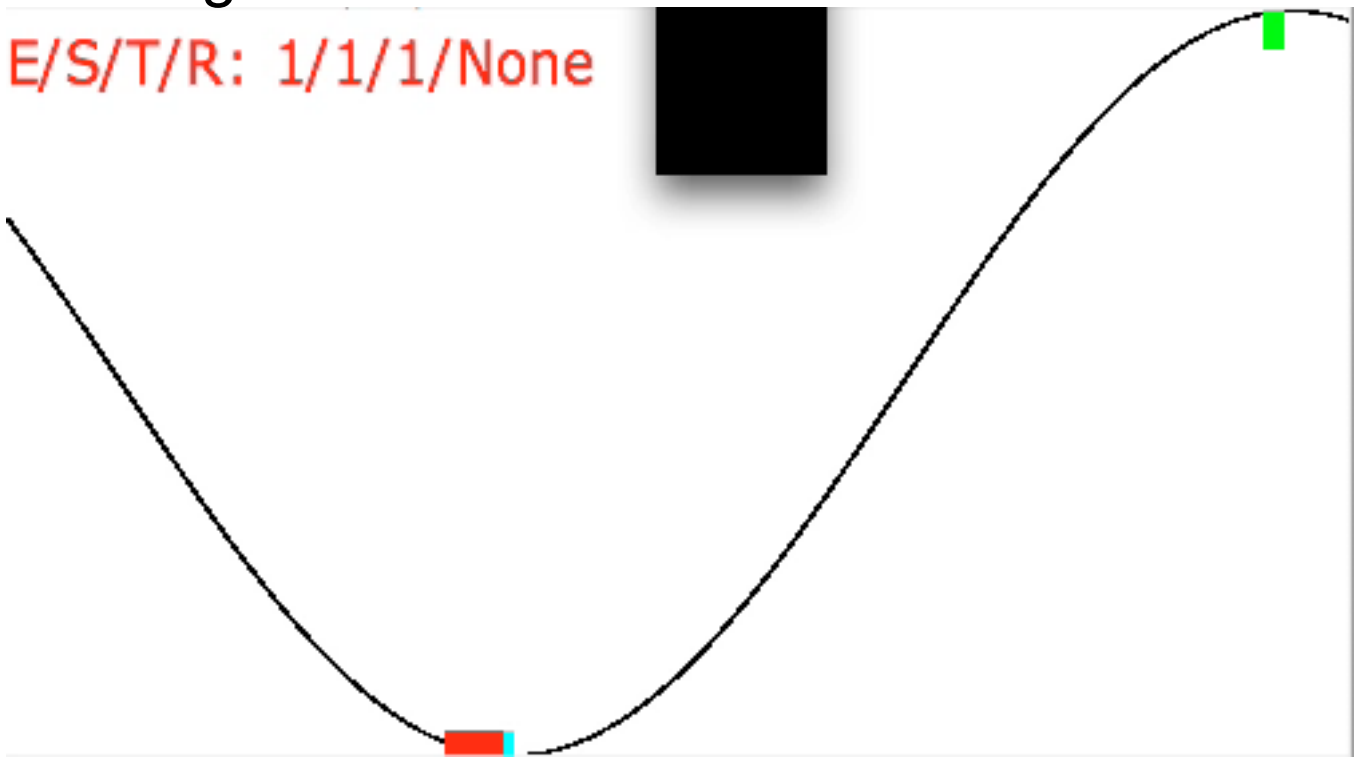
Learning from feedback (interactive shaping)

After training:



Learning from feedback (interactive shaping)

Training:

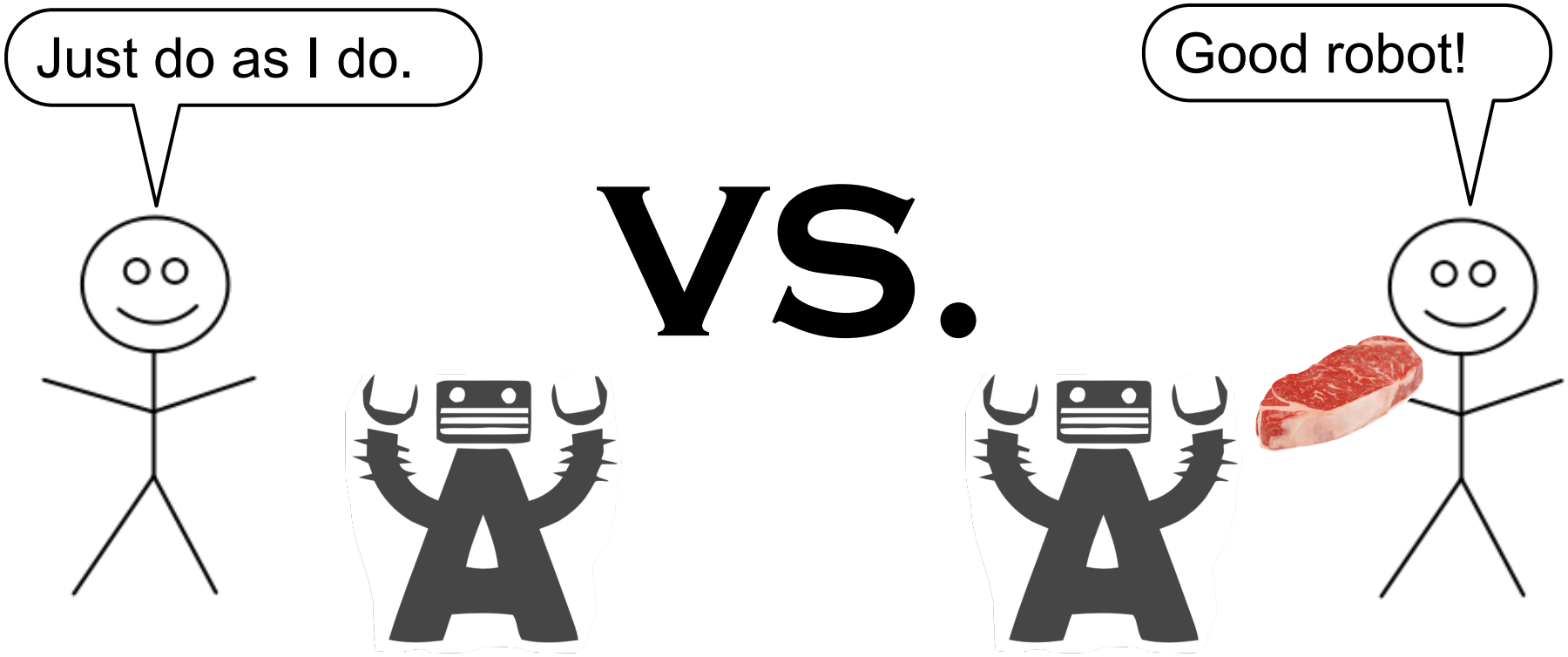


LfD and LfF vs. RL

- Noisy
- Limited by human ability
- Requires human's time

- Faster learning
- Empowers humans to define task

And out come the contendas!!



Learning from Demonstration
(LfD)

Learning from Feedback
(LfF)

An a priori comparison

Demonstration more specifically points to the correct action

Interface

- LfD interface may be familiar to video game players
- LfF interface is simpler and task-independent

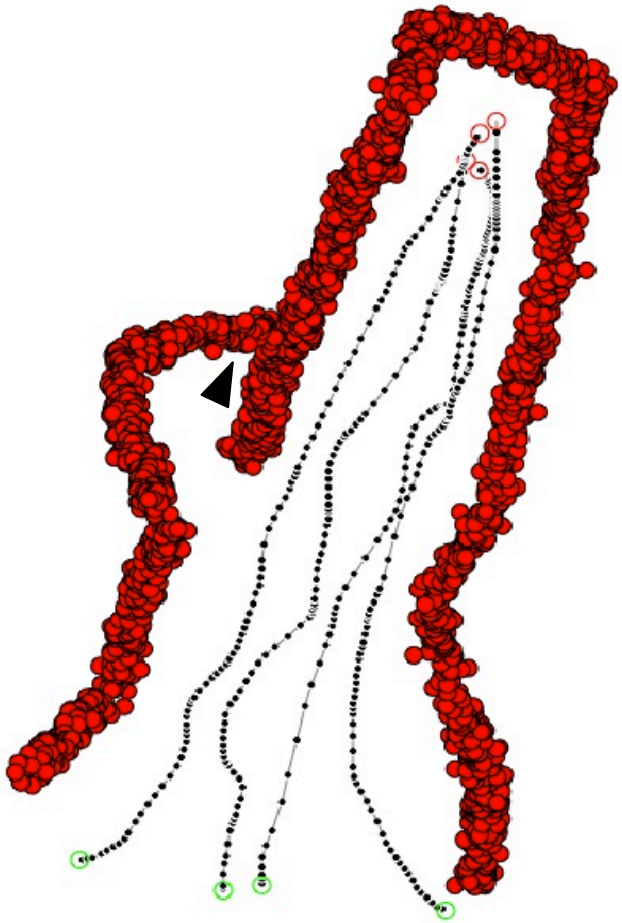
An a priori comparison

Expression of learned model during training:

LfF? yes.

LfD? generally no.

- LfD - better initial training performance
- LfF - can observe and address model's weaknesses
- LfF - training and testing performance match up better



Painted with MLDemos software

An a priori comparison

Task expertise

- LfF - easier to judge than to control
- Easier for human to increase expertise while training with LfD

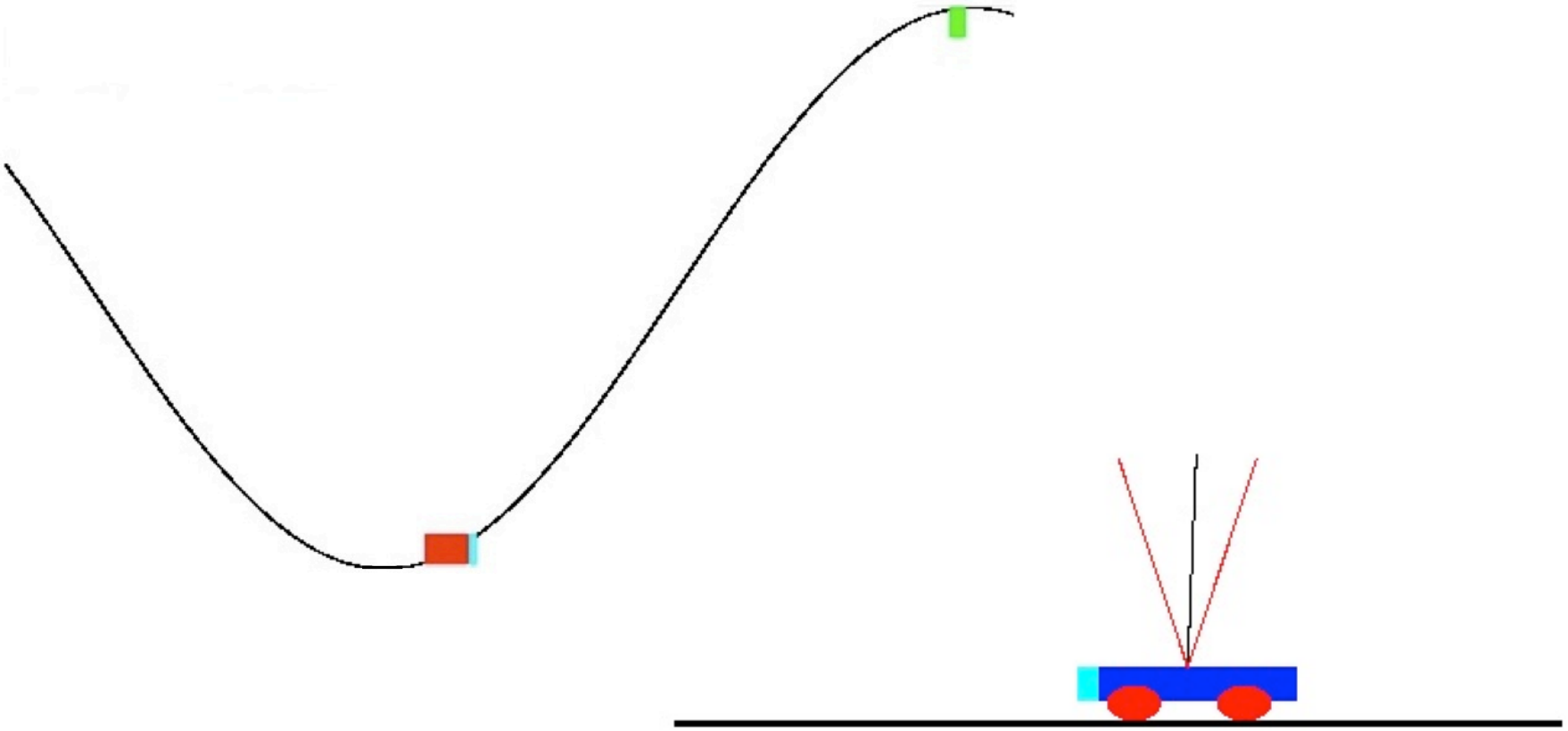
Cognitive load - less for LfF

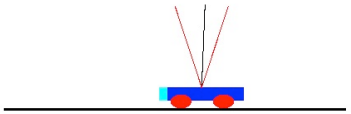
An a priori comparison

General hypothesis

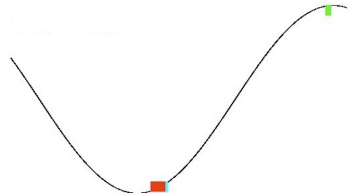
LfD generally performs better,
but situation-dependent

Pilot study





Pilot study



16 undergraduates

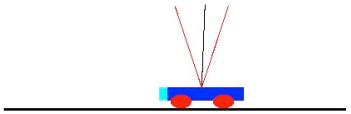
Cart Pole first, then Mountain Car

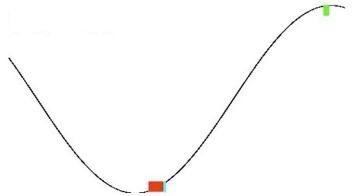
- Practice and test rounds
- Randomized: LfF or LfD first
  - Unbalanced result: LfF was first for 87.5% of CP and 69% of MC

Keyboard interface

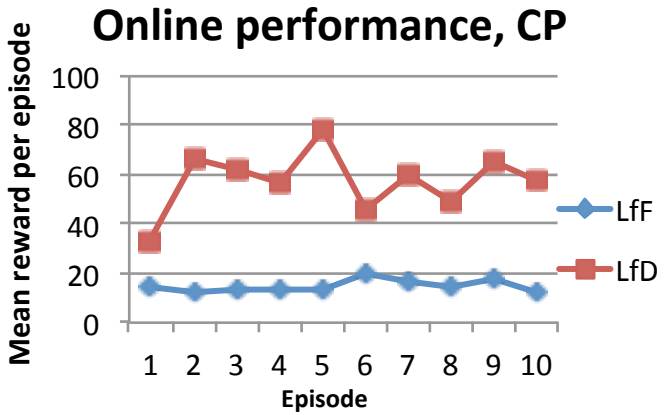
- LfD: j, k, l
- LfF: z, /

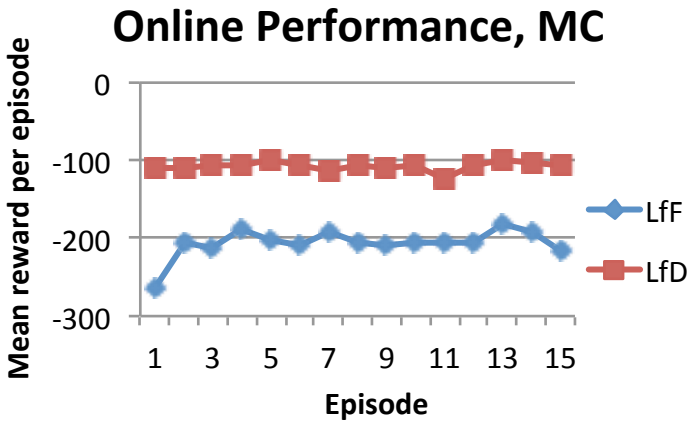
Pilot study

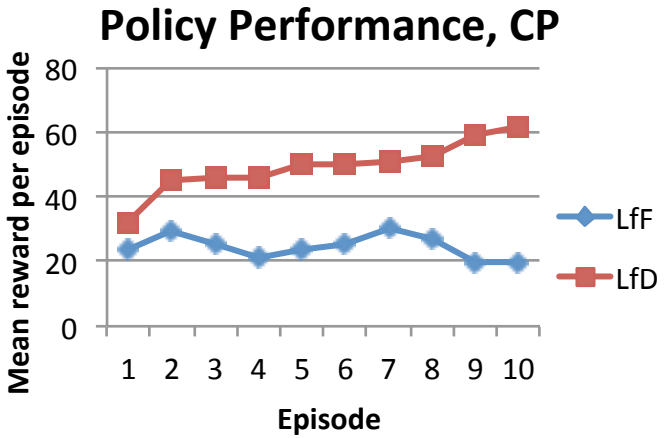


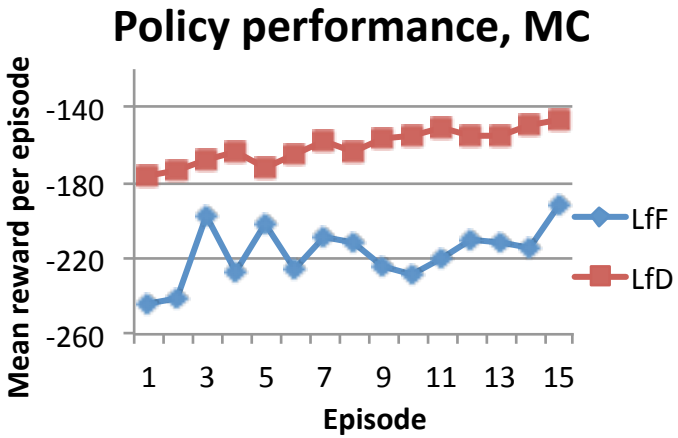


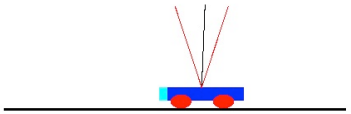
Main result



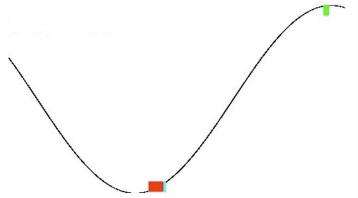




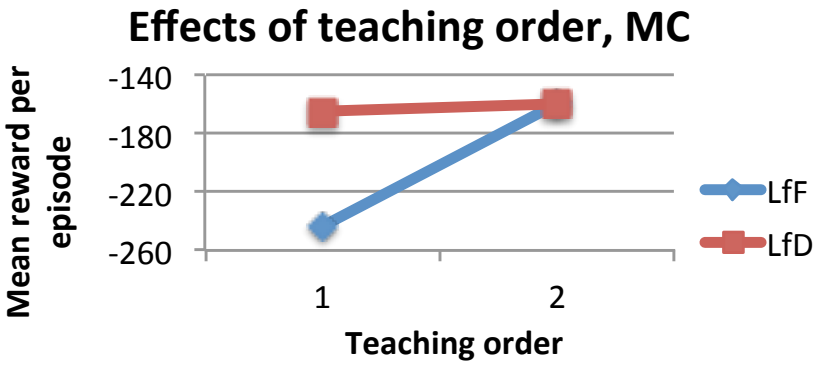


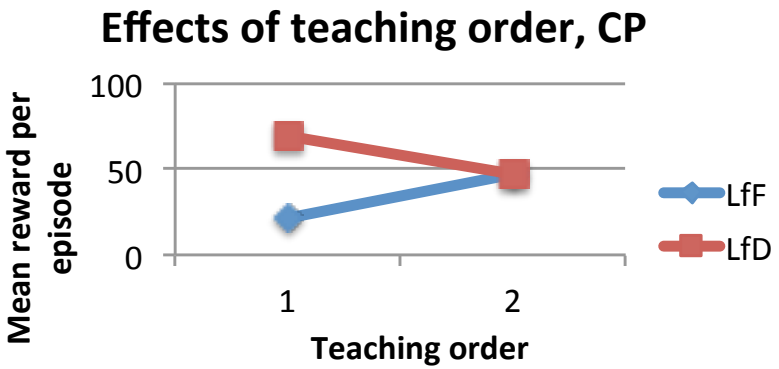


Pilot study

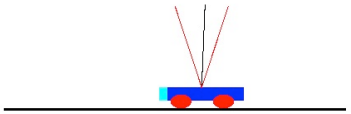


Interaction effects

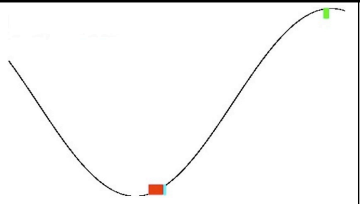


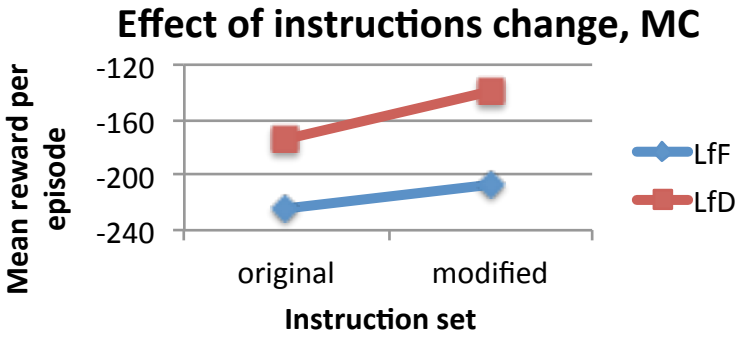


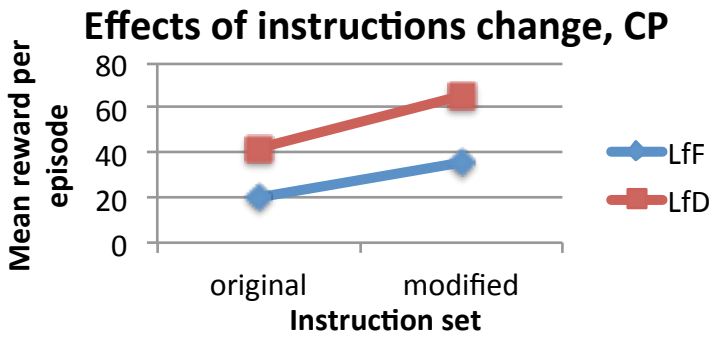
Pilot study



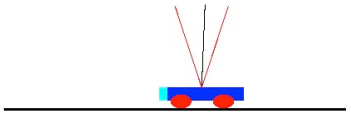
Interaction effects



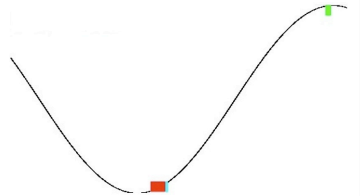




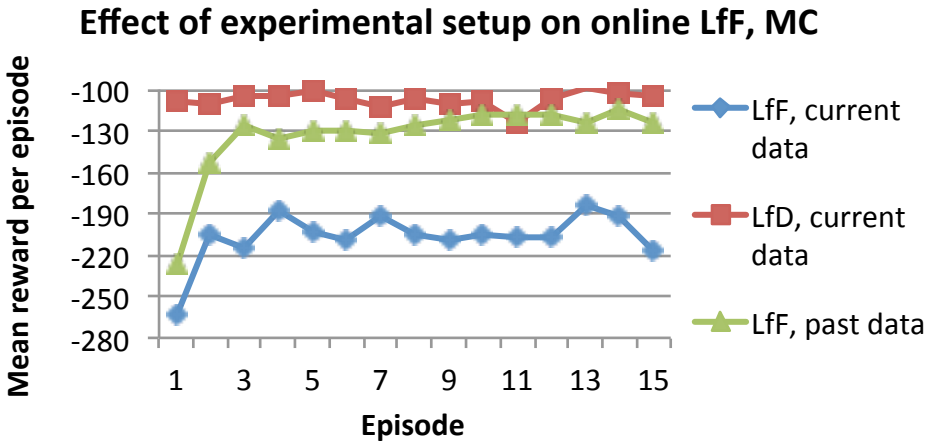
Added a verbal instruction to give frequent feedback for LfF.



Pilot study



Interaction effects

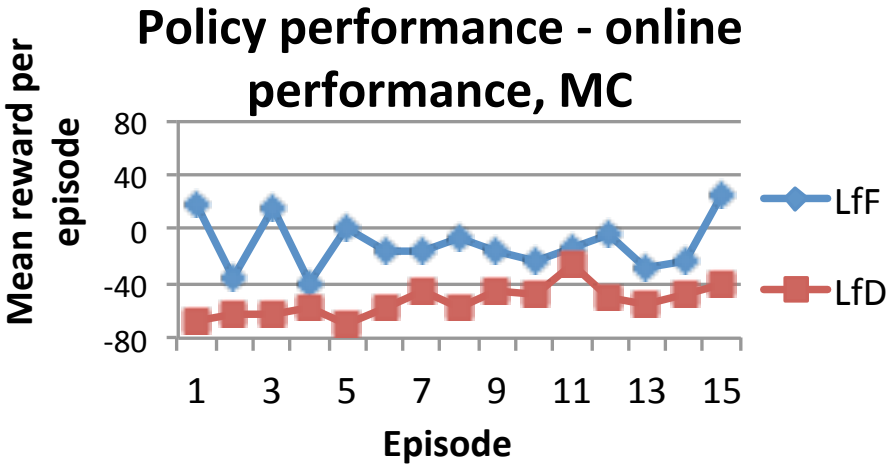


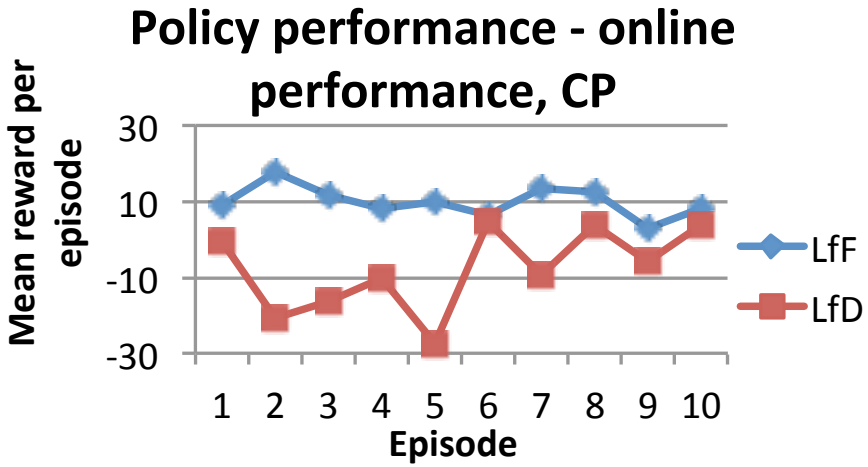
Previous experiment differed:

- more subject preparation
- announced high scores in progress
- ...

Pilot study

Online vs. offline performance





Tentative takeaways from performance comparisons

LfD was better in our experiments.

But both were sensitive to the experimental setup.

Tentative takeaways from performance comparisons

Subjects need more preparation for LfF.

- With *zero* task expertise, LfD still allows learning on the job
- LfF vs. LfD interfaces

Tentative takeaways from performance comparisons

LfD's offline, learned performance is generally worse than its training samples.

LfF's offline, learned performance is generally as good or better than during training.

To conclude,

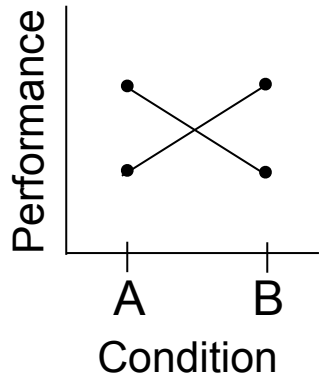
Results

- LfD was better.
- But performance was situational.
- LfF needed more subject preparation.
- LfF models compared better to training performance.

To conclude,

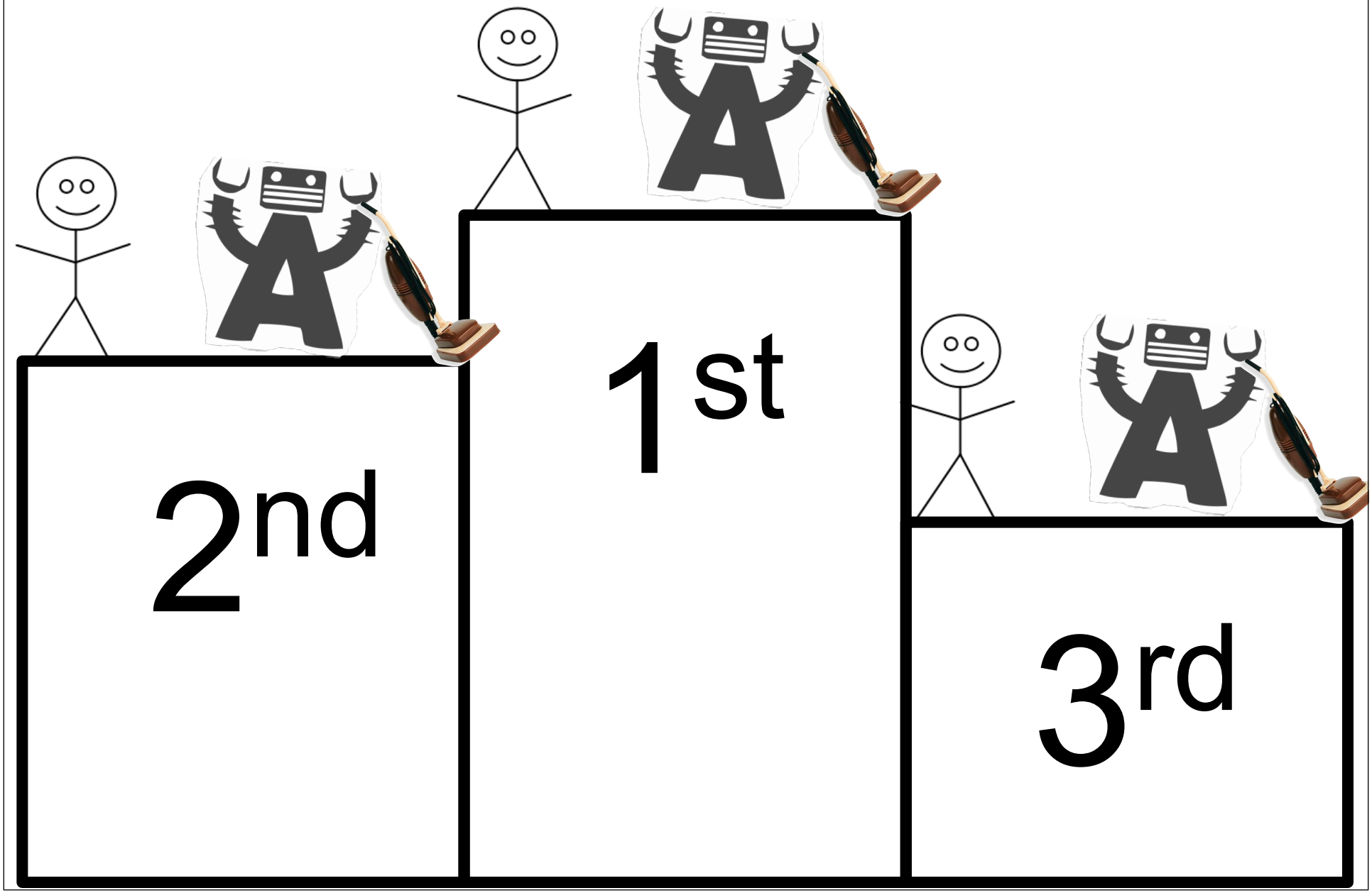
Near future work

- More subjects
- More balanced conditions
- More interesting manipulations (e.g., model representation and control interface quality)
- Aim for crossover interactions

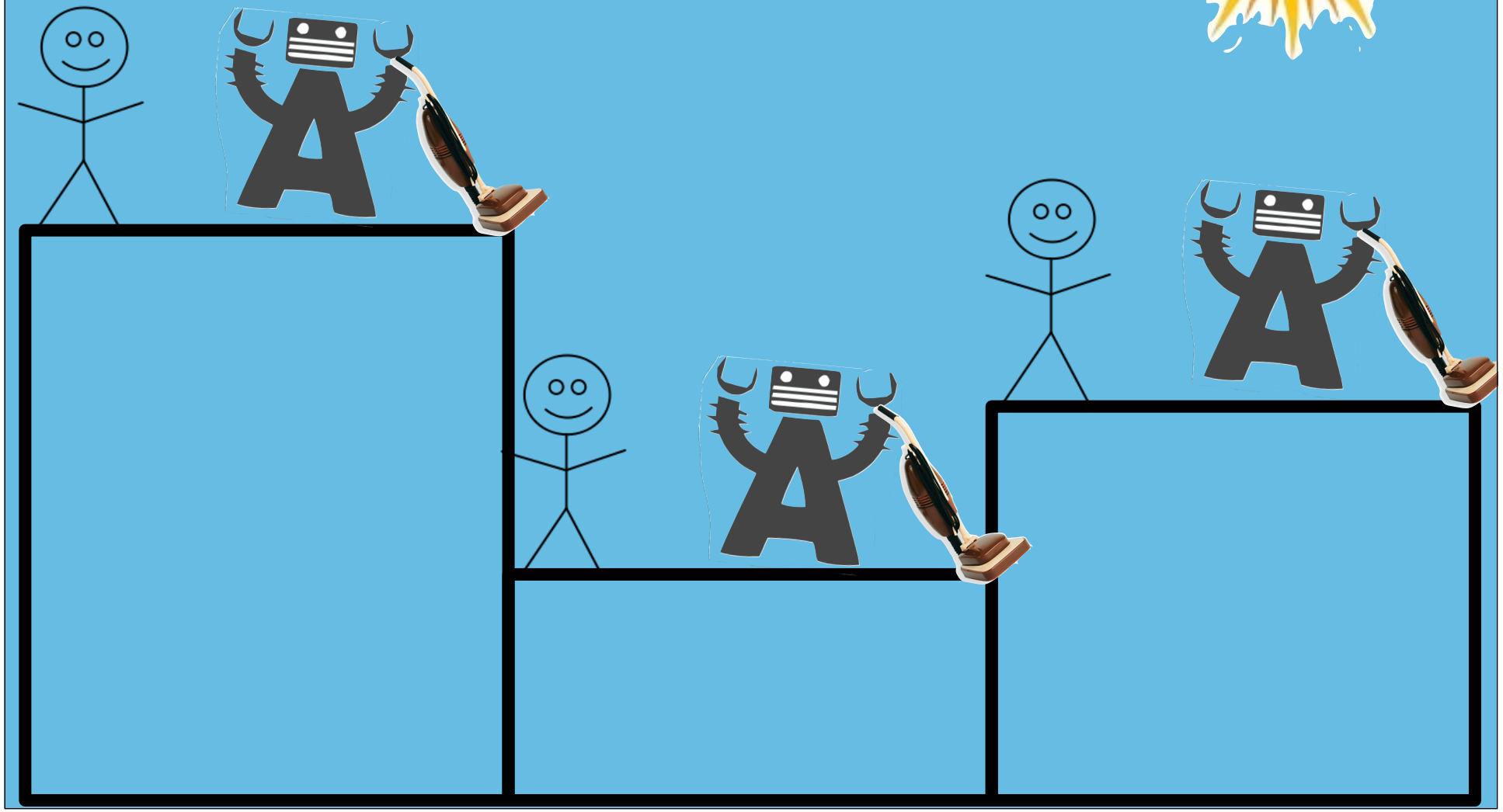


- Learn from both LfD and LfF!

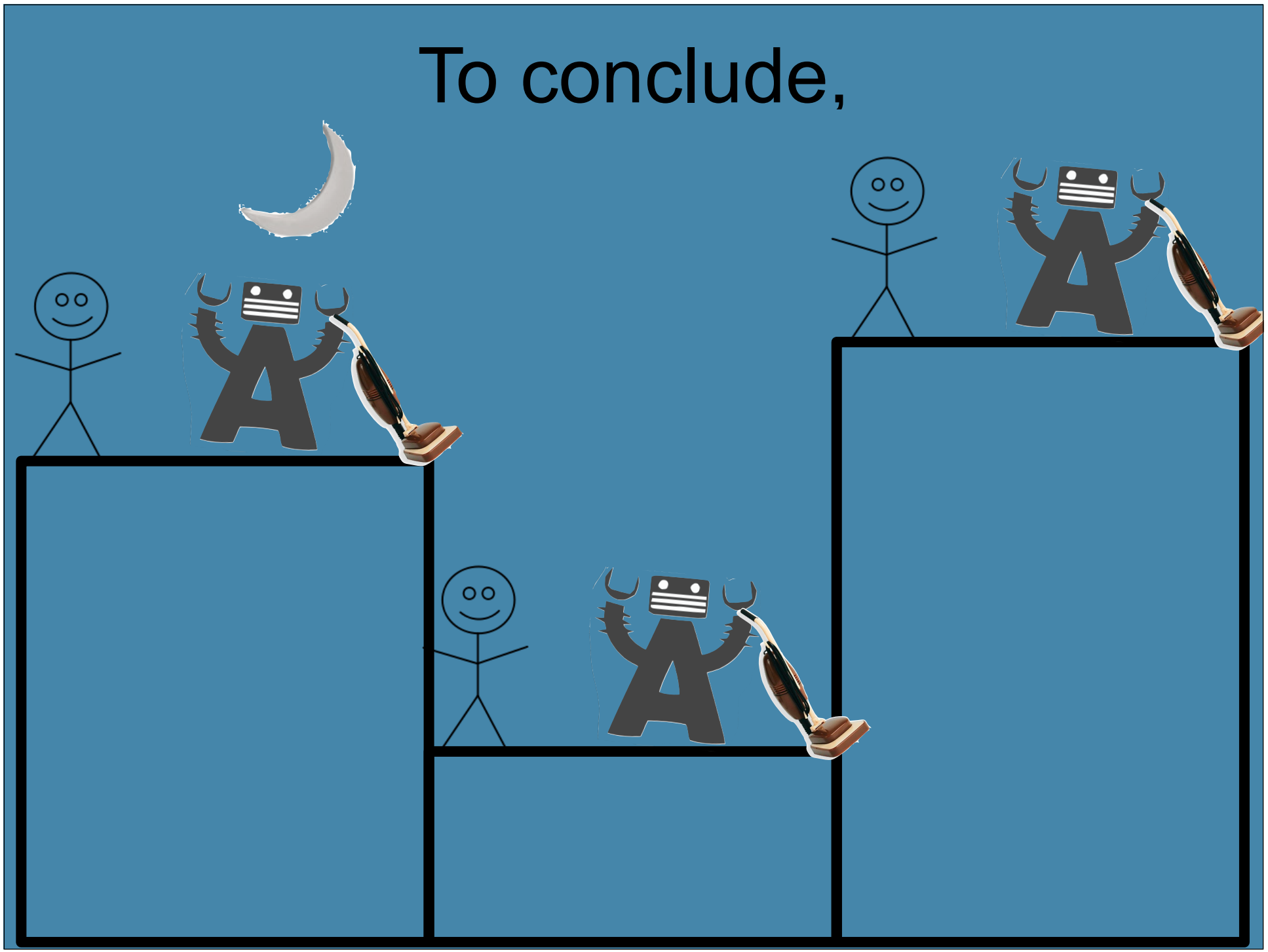
To conclude,



To conclude,



To conclude,



To conclude,