# Reinforcement Learning with Human Teachers: Evidence of Feedback and Guidance with Implications for Learning Performance

Andrea L. Thomaz and Cynthia Breazeal

MIT Media Lab

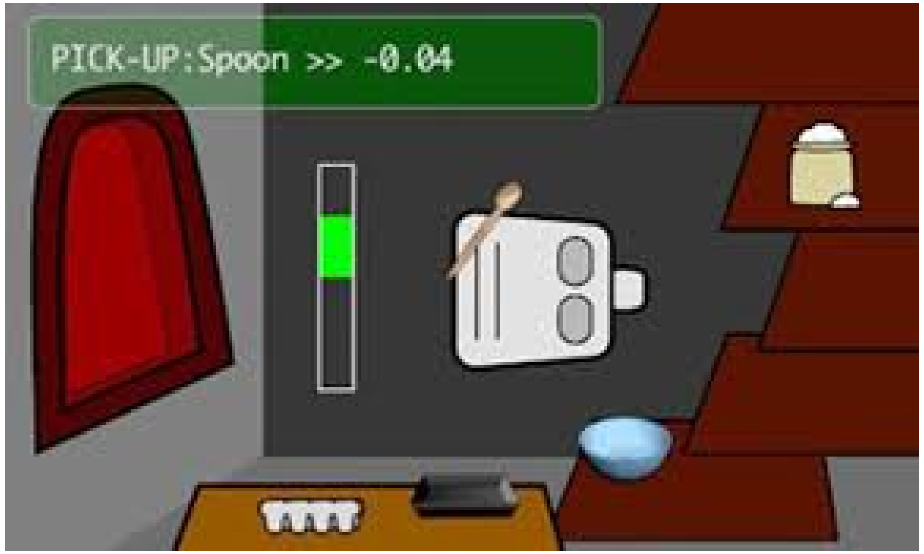
alockerd@media.mit.edu, cynthiab@media.mit.edu

AAAI-2006

# Introduction

- Robert learn by interaction
- Human provide reward signal
- Does people use reward only about past actions?  Or as future guidance?

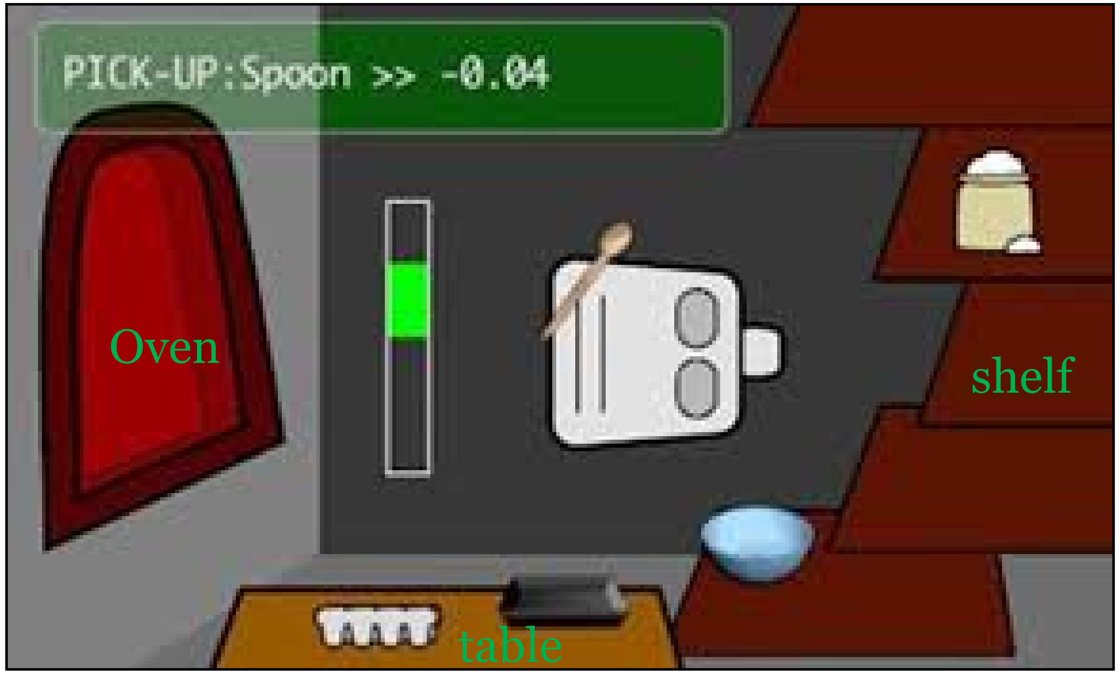
# Experimental Platform: Sophie’s Kitchen




Agent: Sophie
Objects: 面粉, 鸡蛋, 调羹, 碗和 托盘
Goal: 烘焙蛋糕

# MDP

Location set: Shlef, Table, Oven（烤箱）, Agent(the agent in the center surrounded by a shelf, table and oven)




State： agent‘s location, objects’ locations, objects’ state(碗里可能会是空的，只有面粉，只有鸡蛋，两者都有，或者面糊）

Action space: GO left or right, PICK-UP, PUT-DOWN, USE(将面粉倒入碗中）

# Interactive Reward Interface

- Human can provide a reward r = [-1, 1].
- The user receives visual feedback enabling them to tune the reward signal before sending it to the agent
- Can reward the state of a particular object instead the whole state.

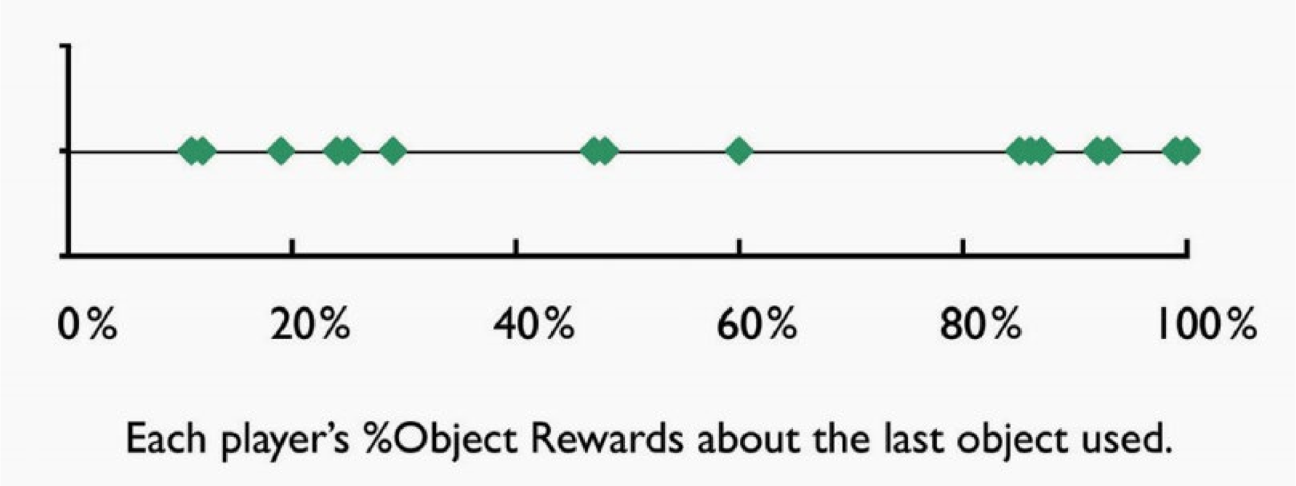
# Experiment

How do human use the reward?

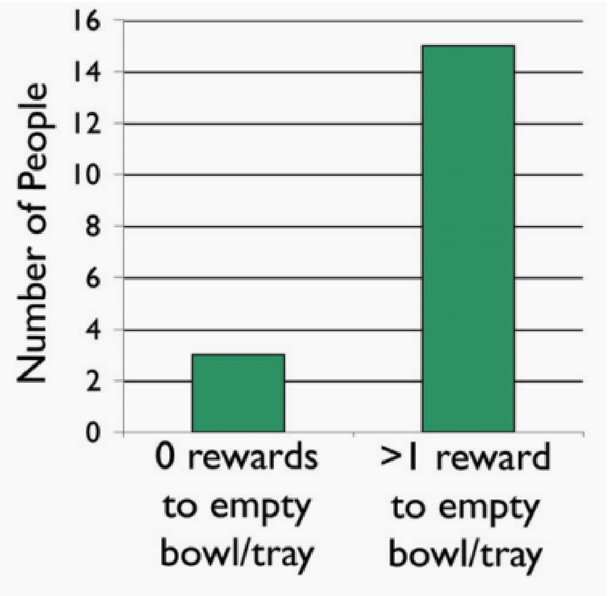
18 paid participants

- Send positive message or negative
- Click on an object, this tells Sophie your message is about that object.

# Result



Each player's %Object Rewards about the last object used.



# Algorithm

**Algorithm 1** Q-Learning with Interactive Rewards:
$s$ = last state, $s'$ = current state, $a$ = last action, $r$ = reward

**while** learning **do**
  $a$ = random select weighted by $Q[s,a]$ values
  execute $a$, and transition to $s'$
  (small delay to allow for human reward)
  sense reward, $r$
  update values:

$$Q[s,a] \leftarrow Q[s,a] + \alpha(r + \gamma(max_{a'}Q[s',a']) - Q[s,a])$$

**end while**

**Algorithm 2** Interactive Q-Learning modified to incorporate interactive human guidance in addition to feedback.

**while** learning **do**
  **while** waiting for guidance **do**
    **if** receive human guidance message **then**
      $g = guide\text{-}object$
    **end if**
  **end while**
  **if** received guidance **then**
    $a$ = random selection of actions containing $g$
  **else**
    $a$ = random selection weighted by $Q[s,a]$ values
  **end if**
  execute $a$, and transition to $s'$
  (small delay to allow for human reward)
  sense reward, $r$
  update values:

$$Q[s,a] \leftarrow Q[s,a] + \alpha(r + \gamma(max_{a'}Q[s',a']) - Q[s,a])$$

**end while**

# Experiments-Expert Data

1. No guidance: feedback only and the trainer gives reward after every action.
2. Guidance: has both guidance and feedback available;

Table 1: An **expert** user trained 20 agents, following a strict best-case protocol; yielding theoretical best-case learning effects of guidance. (F = failures, G = first success).

| Measure | Mean no guide | Mean guide | chg | t(18) | p |
|---|---|---|---|---|---|
| # trials | 6.4 | 4.5 | 30% | 2.48 | .01 |
| # actions | 151.5 | 92.6 | 39% | 4.9 | $<.01$ |
| # F | 4.4 | 2.3 | 48% | 2.65 | $<.01$ |
| # F before G | 4.2 | 2.3 | 45% | 2.37 | .01 |
| # states | 43.5 | 25.9 | 40% | 6.27 | $<.01$ |

# Experiments-Non-Expert Data

*'You can direct Sophie's attention to particular objects with guidance messages. Click the right mouse button to make a yellow square, and use it to help guide Sophie to objects, as in 'Pay attention to this!'*

Table 2: **Non-expert** players trained Sophie with and without guidance communication and also show the positive learning effects of guidance. (F = failures, G = first success).

| **Measure** | **Mean no guide** | **Mean guide** | **chg** | **t(26)** | **p** |
|---|---|---|---|---|---|
| # trials | 28.52 | 14.6 | 49% | 2.68 | <.01 |
| # actions | 816.44 | 368 | 55% | 2.91 | <.01 |
| # F | 18.89 | 11.8 | 38% | 2.61 | <.01 |
| # F before G | 18.7 | 11 | 41% | 2.82 | <.01 |
| # states | 124.44 | 62.7 | 50% | 5.64 | <.001 |
| % good states | 60.3 | 72.4 | | -5.02 | <.001 |

# Conclusion

- People try to guide a agent given reward channel.
- We modify the algorithm to include channel of guidance.
- It improves several dimension of learning
  - the speed of  learning,
  - the efficiency of state exploration,
  - and a significant drop in the number of failed trials