Learning from human-generated reward

(Slides at **tinyurl.com/knoxthesis**)

W. Bradley Knox

Learning Agents Research Group
The University of Texas at Austin

What do I mean by human-generated reward?

Communications of

- approval or disapproval,
- judgments of good or bad behavior or outcomes,
- intention of reward or punishment,
- or something similar

that can be intuitively mapped to a real-valued signal.

Human reward is abundant.

Teaching with human reward

Benefits:

(1) *For undefined tasks*, end users can specify correct behavior.


Image is courtesy of ABB

(2) *For defined tasks*, human task knowledge can be transferred to aid learning.

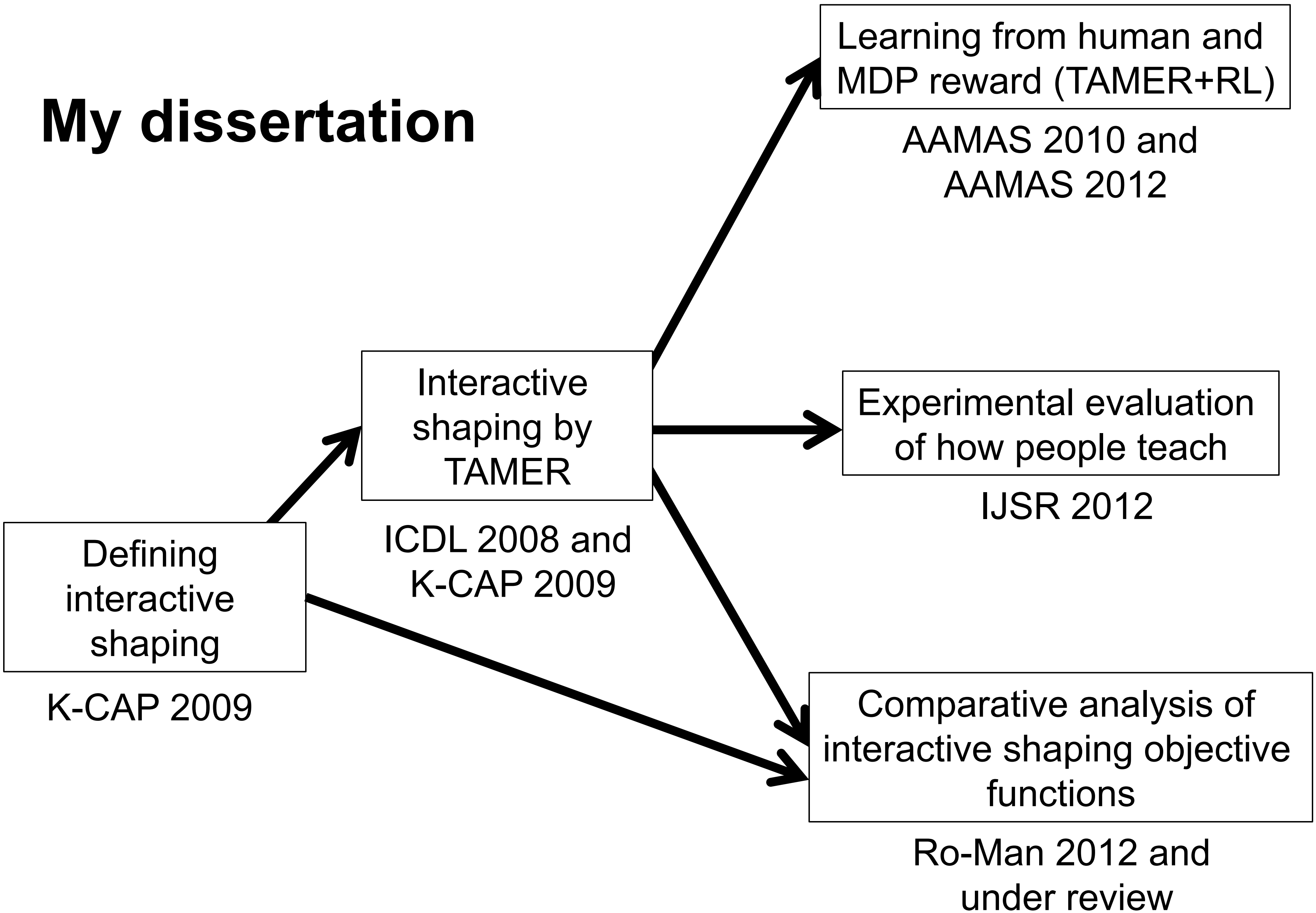
… without requiring programming skills.

Research Question:

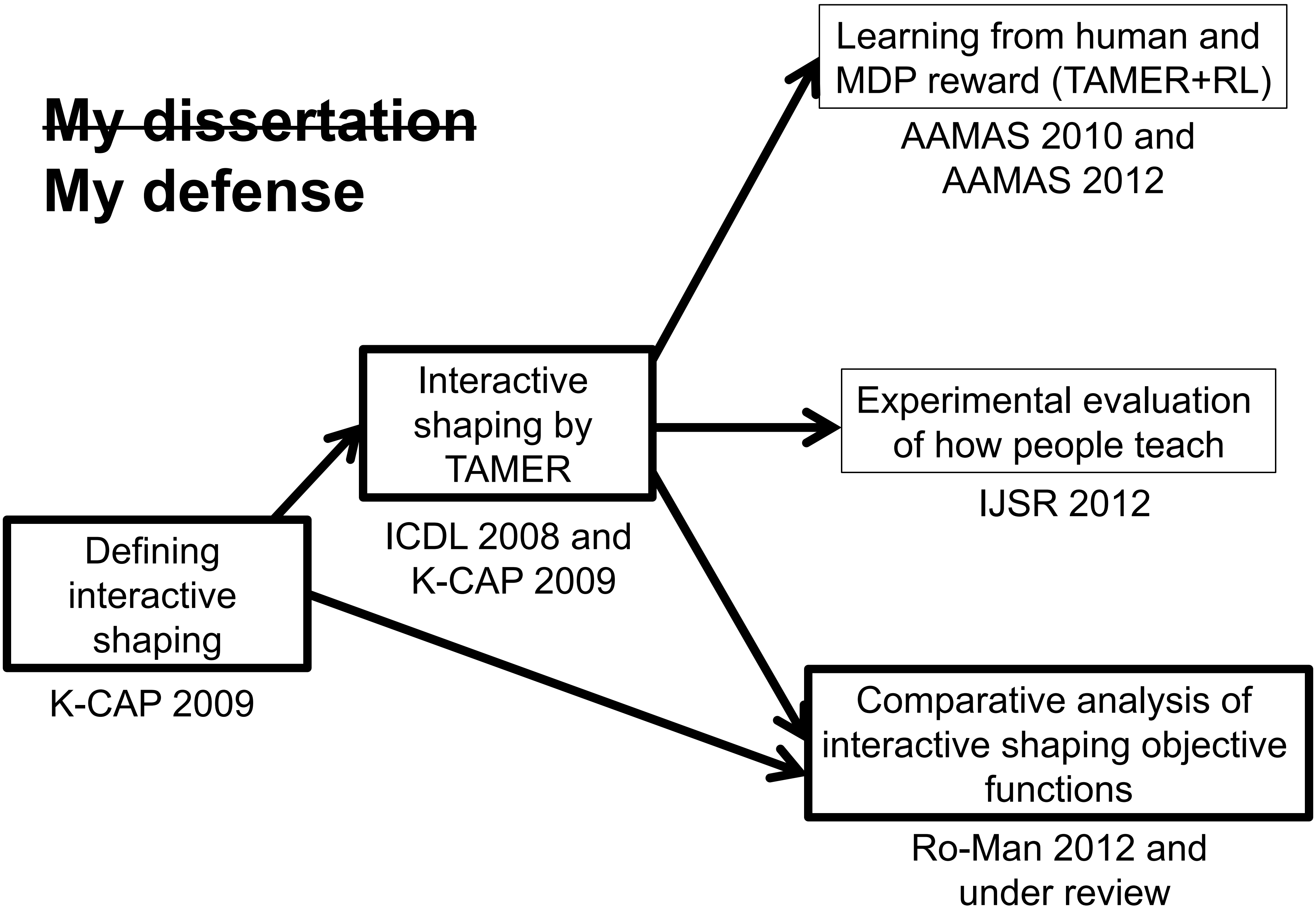
How can agents harness
the information contained in
human-generated signals of **reward***
to learn **sequential decision-making tasks**?

*Includes both positive and negative values

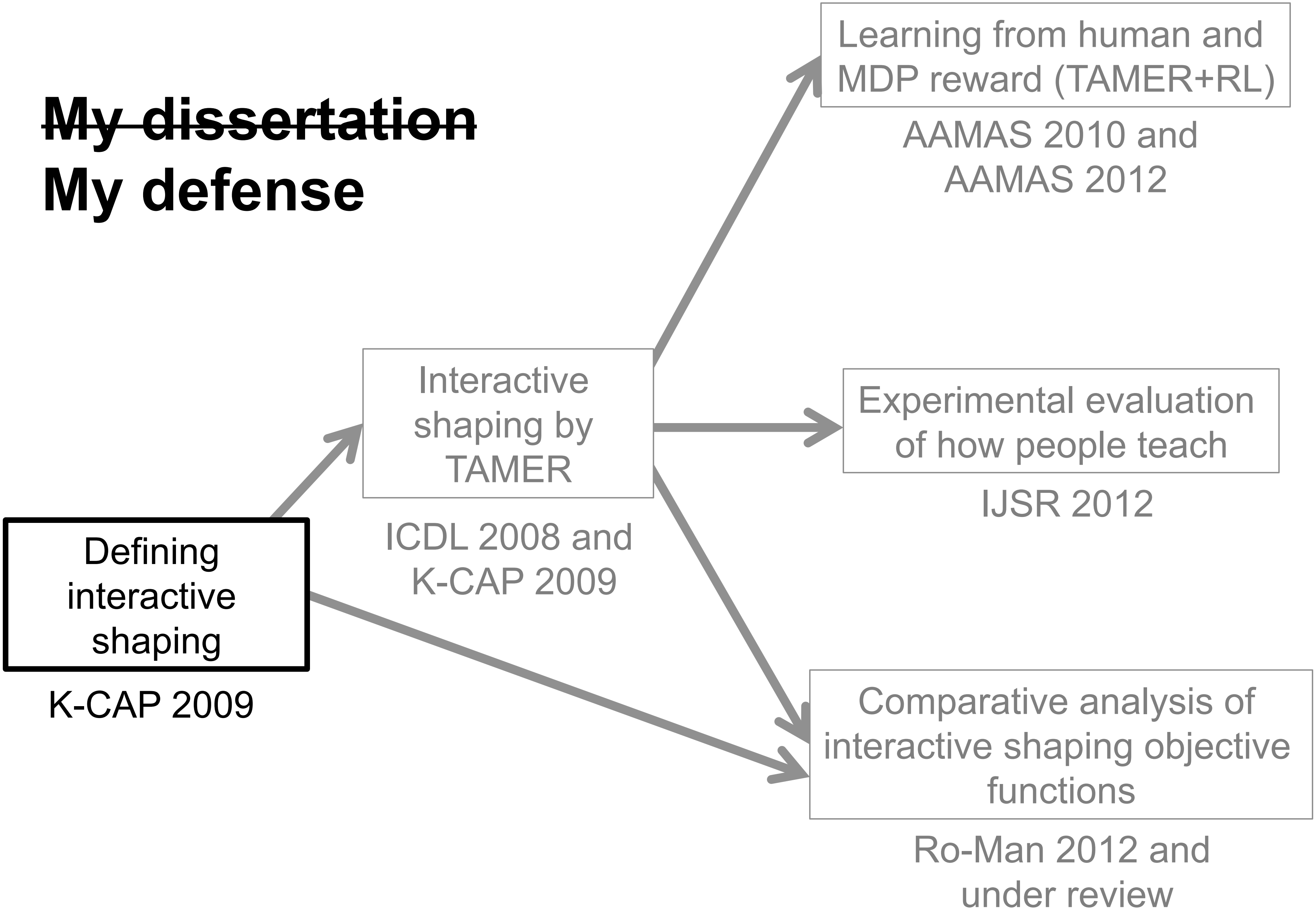
My dissertation



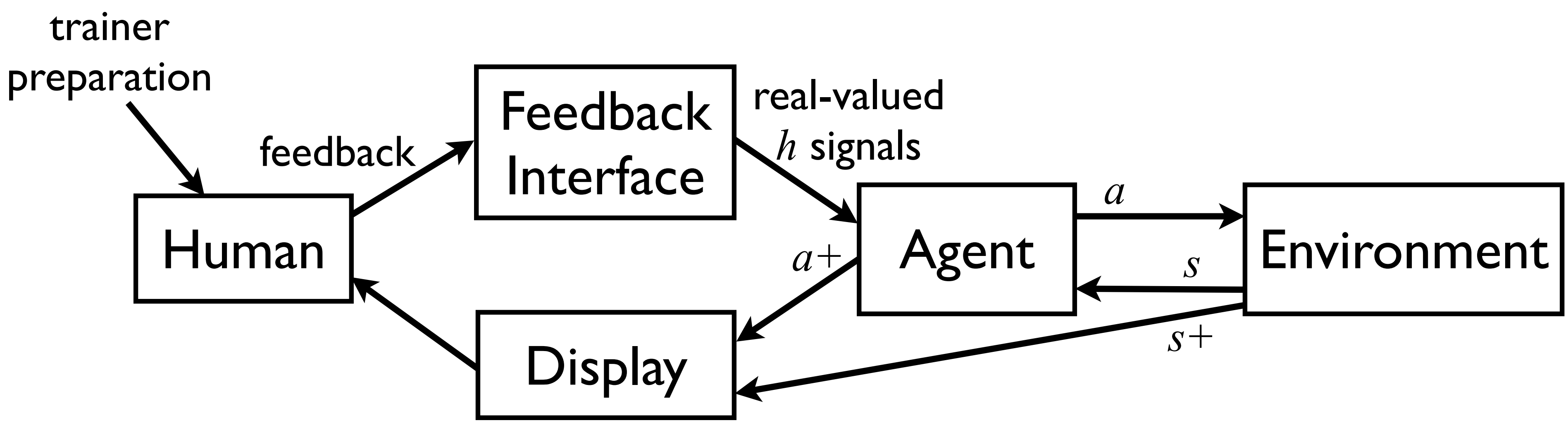
~~My dissertation~~
My defense



~~My dissertation~~
My defense



1 Information flow in interactive shaping



1 The Interactive Shaping Problem

Human trainer:

- observes agent
- delivers reward signals $h_i \in \mathbb{R}$ at any time
- attempts to maximize task performance by τ

Each time step, agent:

- receives state description $s \in S$
- chooses an action $a \in A$

1 The Interactive Shaping Problem

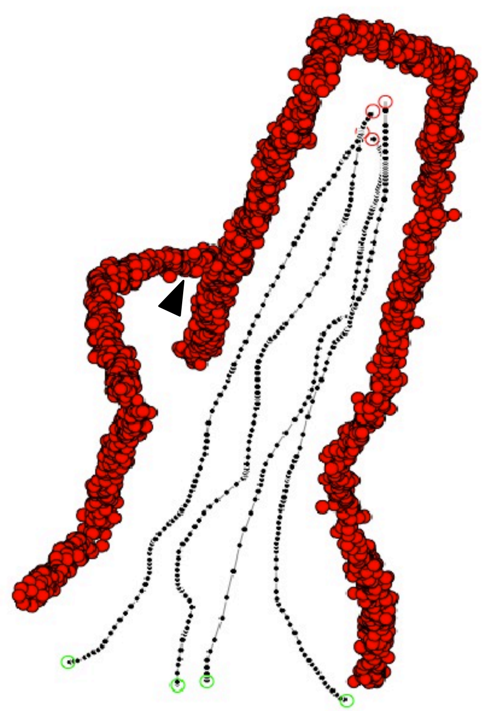
Given this input,

1. define the agent's objective with respect to reward received such that it maximizes task performance (by τ), and
2. optimize with respect to the objective

Interactive shaping vs. learning from demonstration

Advantages of interactive shaping

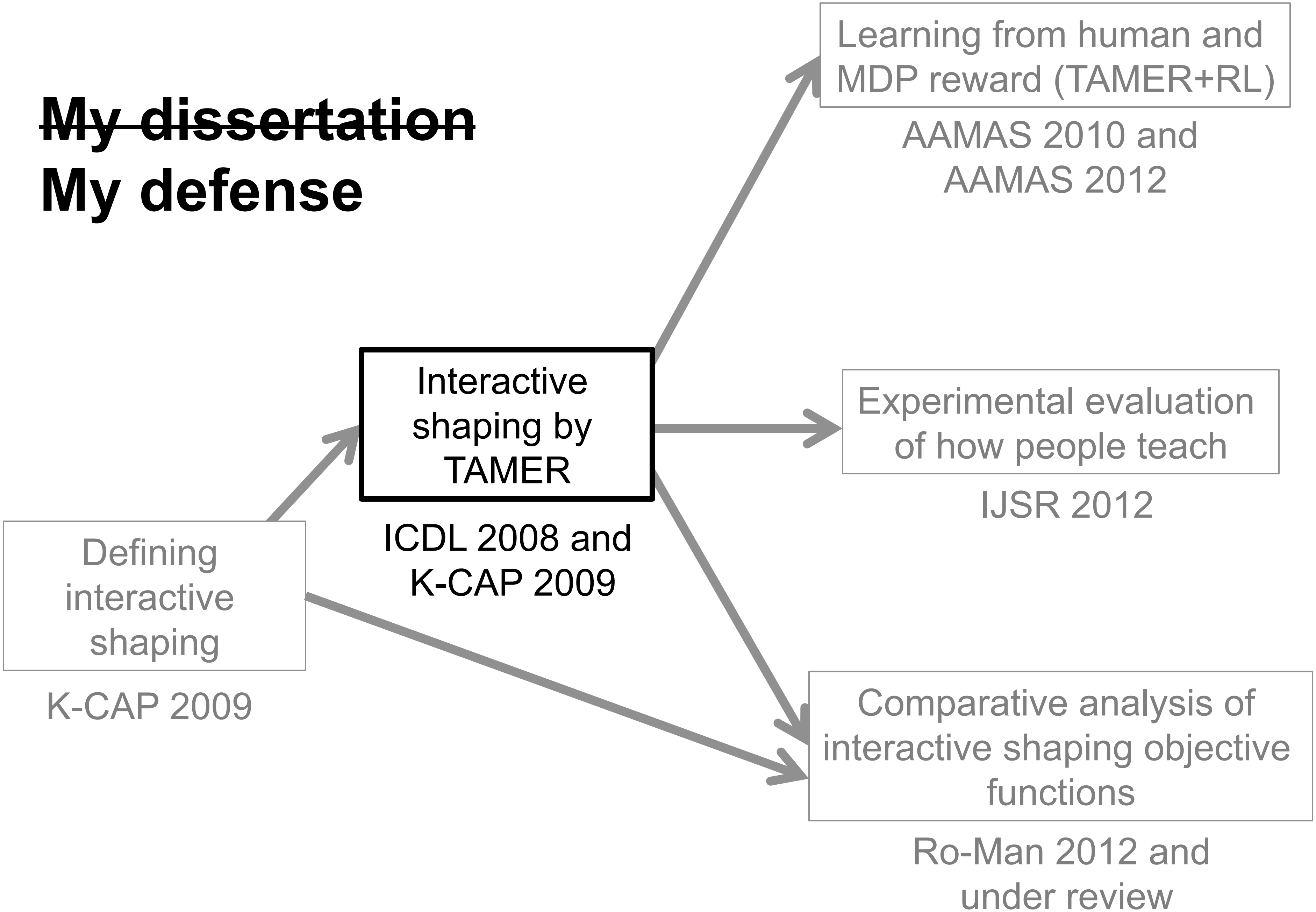
- yields information on the policy actually learned
- criticism requires less expertise than action
- task can be specified, not just policy
- cognitive load
- agent-independent interface



Painted with MLDemos software

But demonstration does allow a policy to be directly specified.

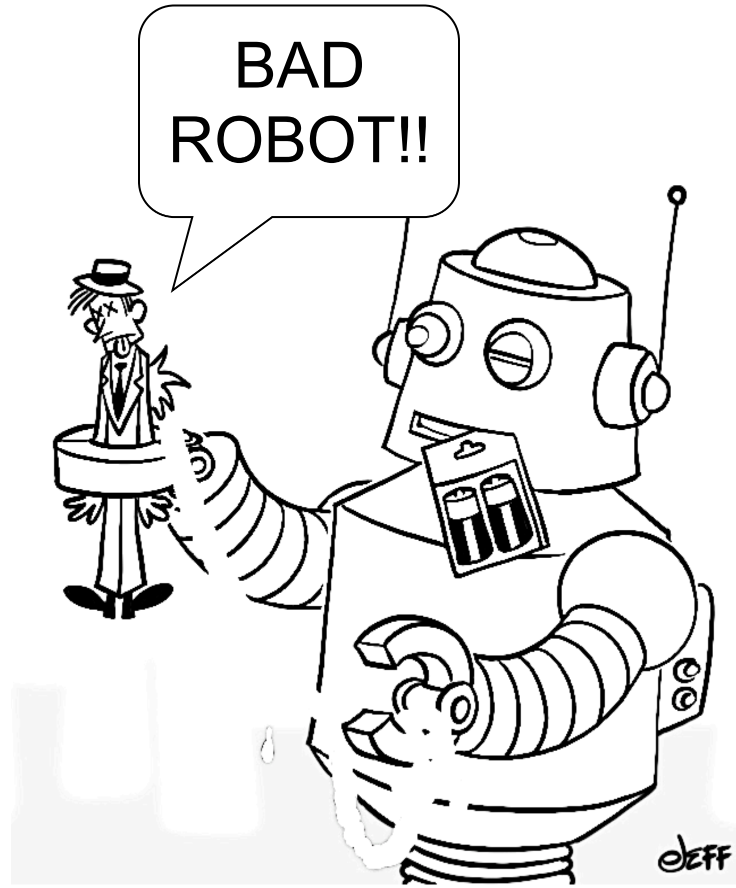
~~My dissertation~~
My defense



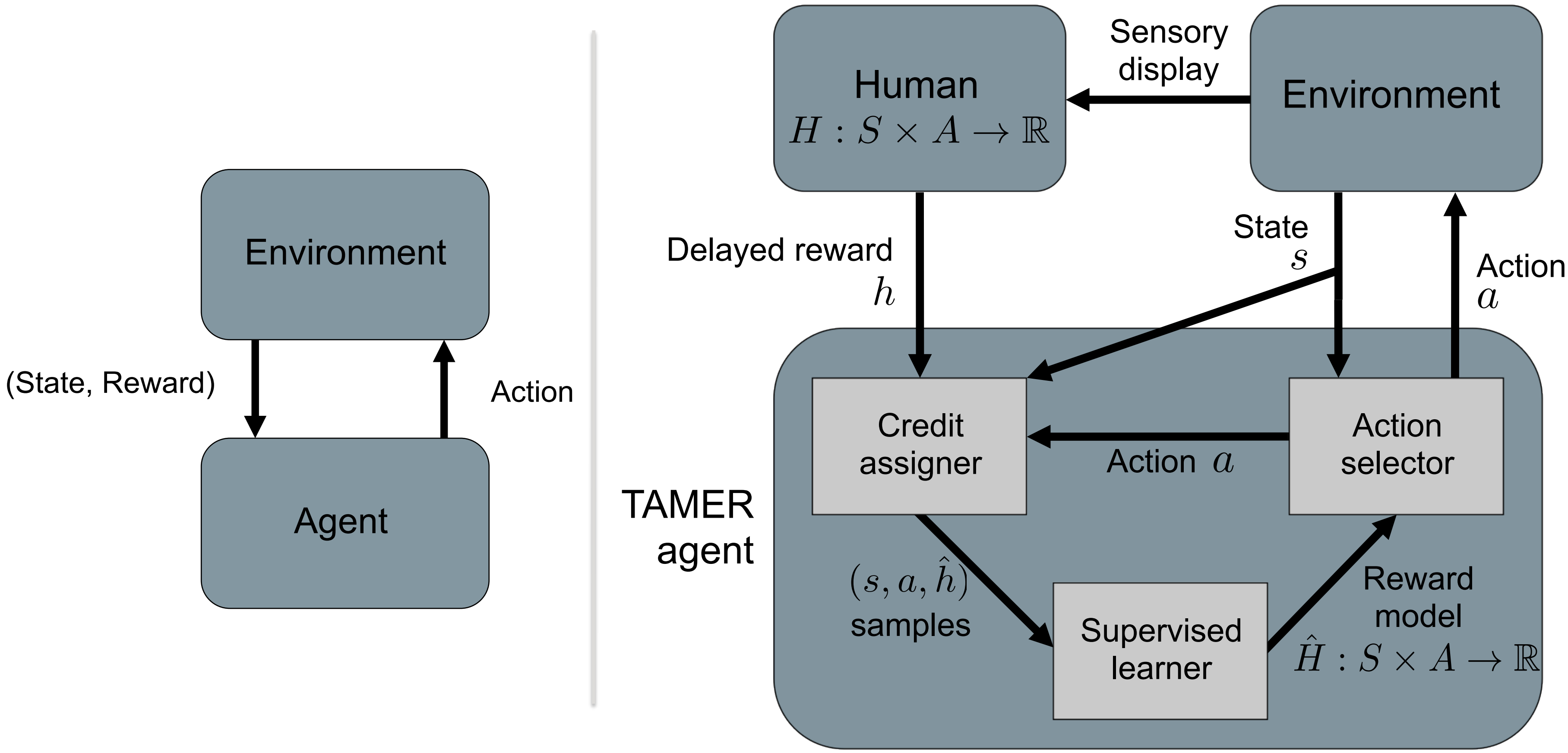
2 One solution to interactive shaping

Two insights:

- Trainer has long-term impact in mind.
  - We can consider reward a full judgment of desirability of behavior.
- Trainer can reward with small delay.



Teaching an Agent Manually via Evaluative Reinforcement (**TAMER**)



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

ICDL 2008 and K-CAP 2009

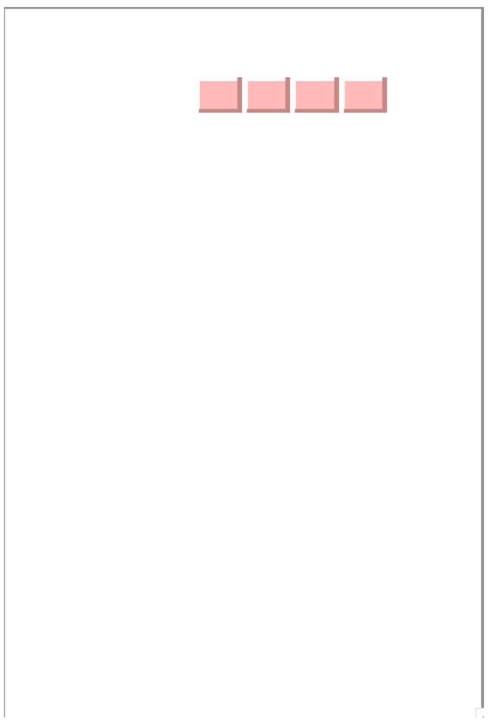
Teaching an Agent Manually via Evaluative Reinforcement (**TAMER**)

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

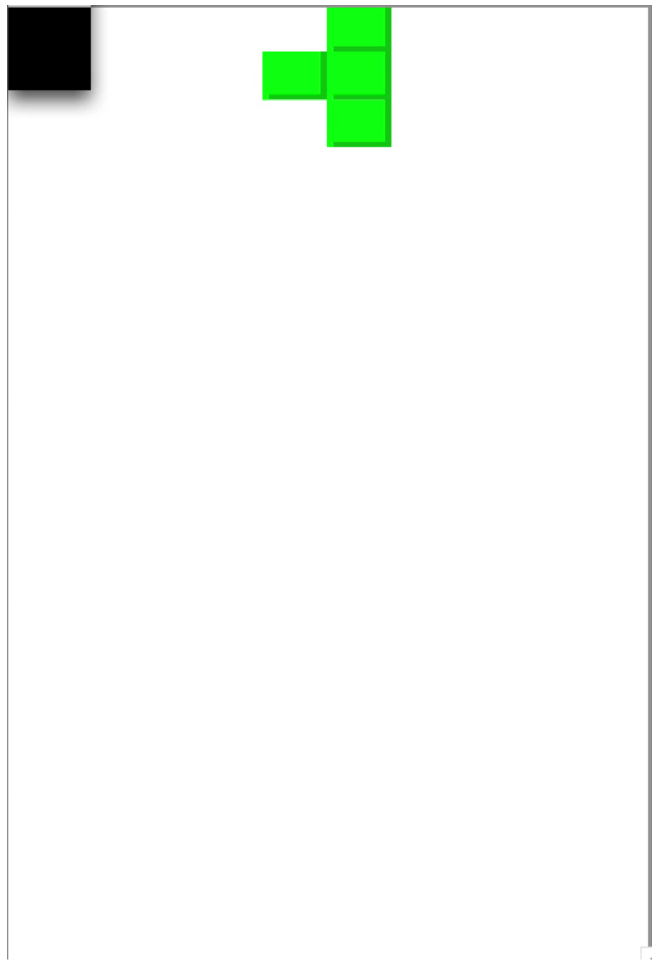
I.e., TAMER **reduces** an apparent reinforcement learning problem **to a supervised learning problem** by setting $\gamma=0$.

TAMER in action: Tetris

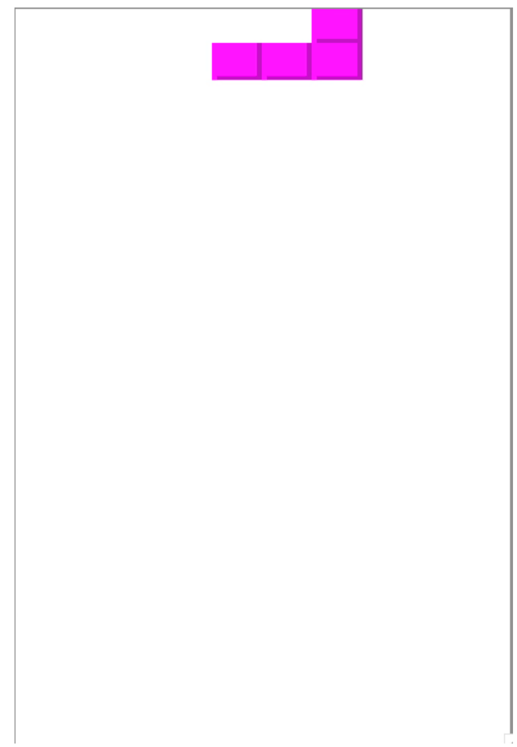
Before training:



Training:

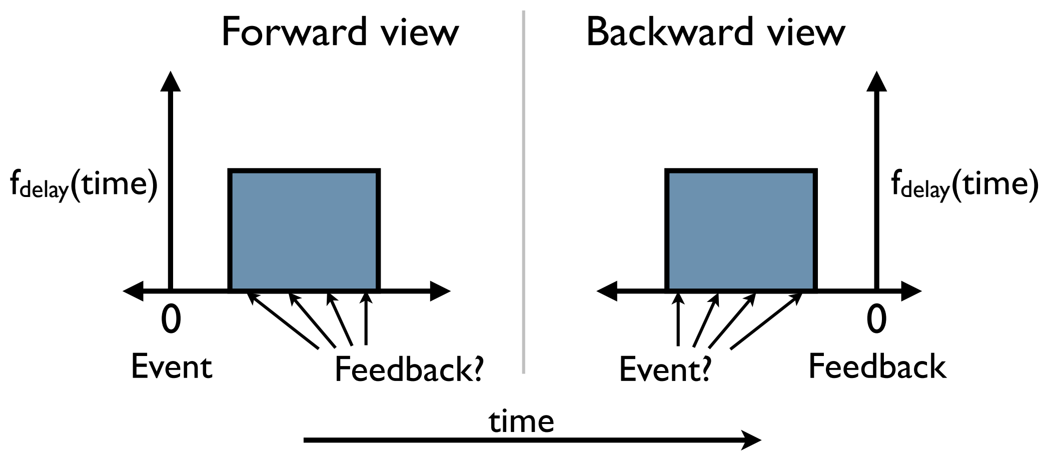


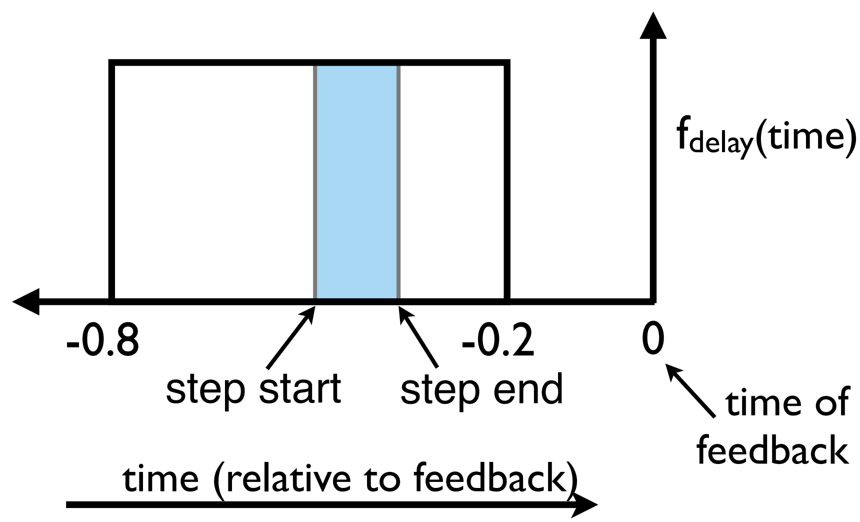
After training:



Environment courtesy of RL-Library and RL-Glue

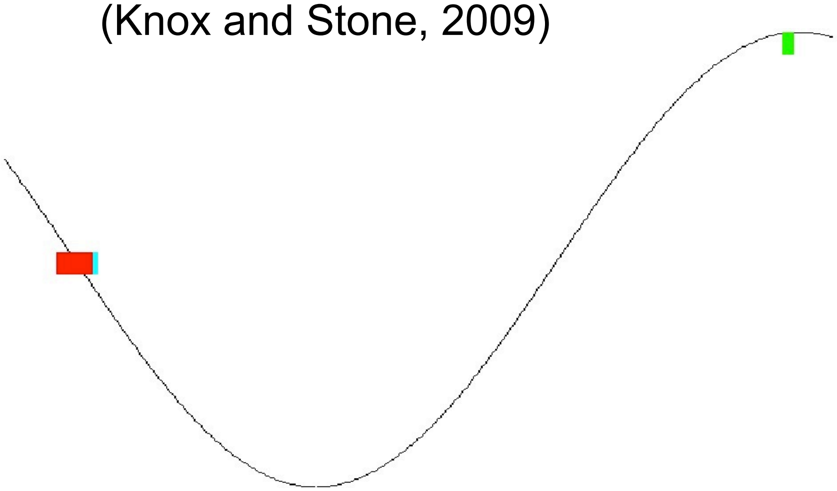
Handling reward delay



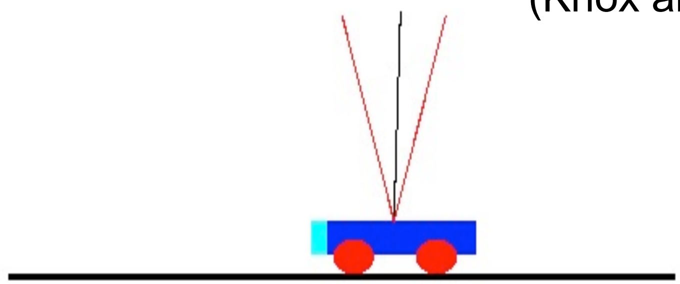


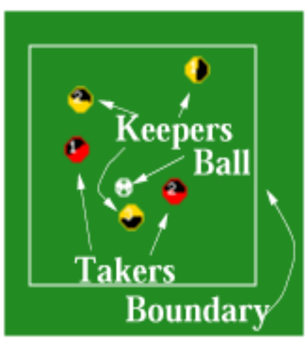
TAMER success on other domains

Mountain Car
(Knox and Stone, 2009)

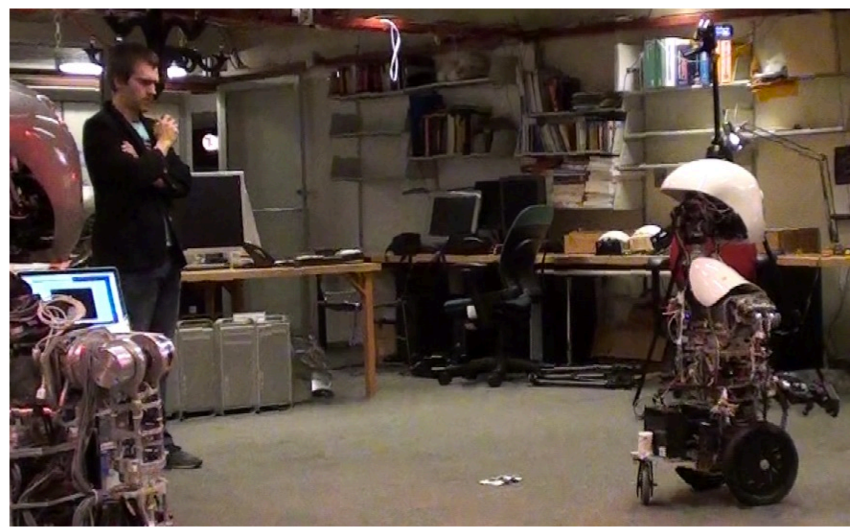


Balancing Cart Pole
(Knox and Stone, 2012)





3 vs 2 Keepaway
(Sridharan, 2011)

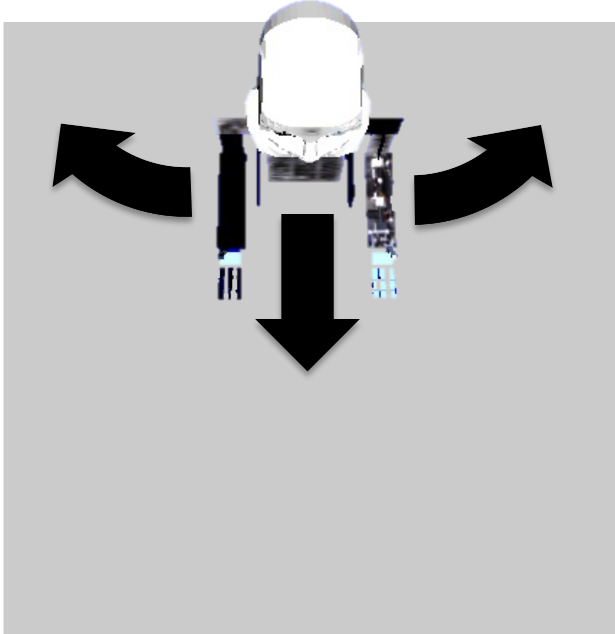


Interactive robot navigation
(Knox, Stone, and Breazeal, 2012)

Environments courtesy of RL-Library and RL-Glue (adapted)

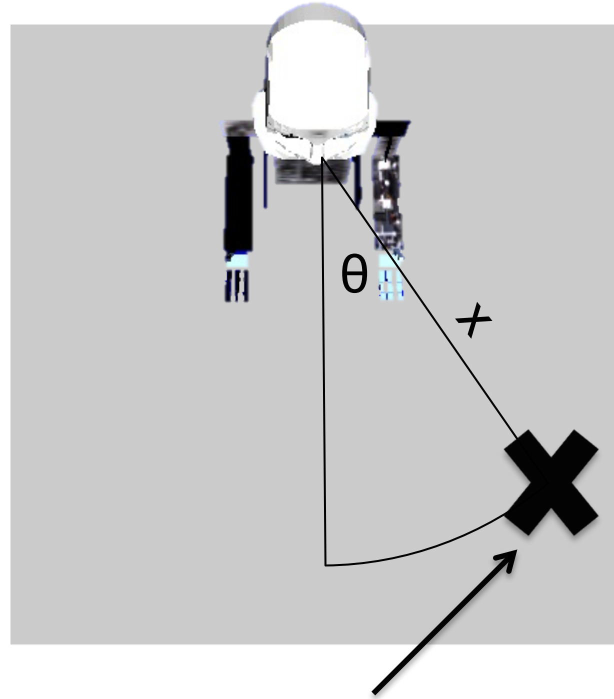
TAMER in action: interactive robotics

4 actions:



and “stay”

2 state features:



training artifact

Reward interface:

Knox, Stone, and Breazeal, 2012

TAMER in action: interactive robotics

Knox, Stone, and Breazeal, 2012

TAMER in action: interactive robotics

Knox, Stone, and Breazeal, 2012

TAMER Results

When compared to human-less algorithms learning from predefined "MDP reward" functions:

TAMER learns with fewer samples

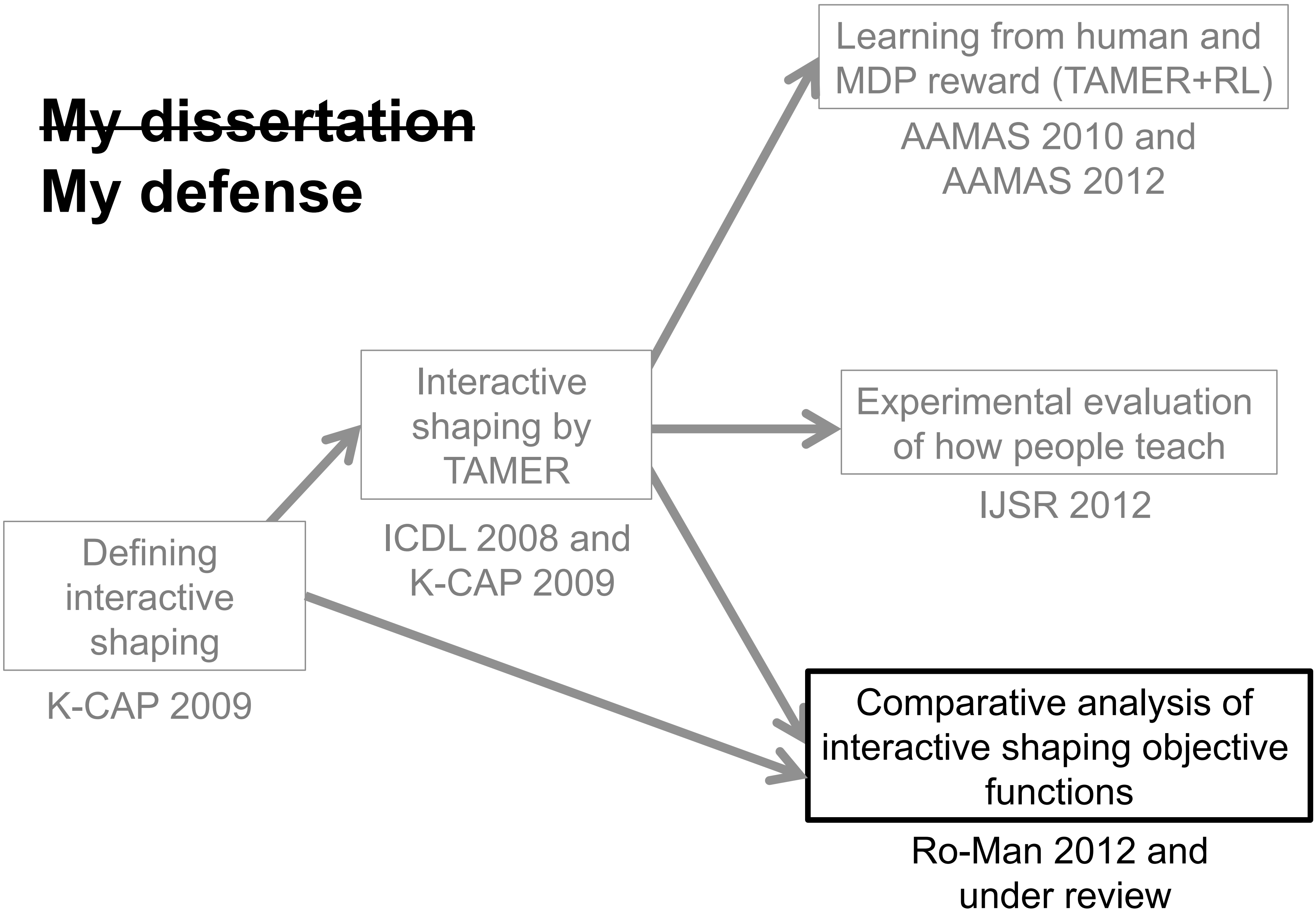
and

learners using MDP reward eventually equal or surpass TAMER

Interactive shaping solutions

Earliest publication	**Number of tested domains**	**Reward interface**	**Effective discount factor**	**Addresses reward delay?**	**Models human reward?**
Isbell et al. (2000)	1	Typed words	0.7	Implicitly	No
Thomaz and Breazeal (2006)	1	Mouse gestures	0.75	Implicitly	No
Knox et al. (2008) - TAMER	**6**	**Push buttons**	**0**	**Explicitly**	**Yes**
Tenorio-Gonzalez et al. (2010)	2	Verbalized words	0.9	Implicitly	No
Suay and Chernova (2011)	1	Mouse gestures	0.75	Implicitly	No
Pilarski et al. (2011)	1	Push buttons	0.99	Implicitly	No

~~My dissertation~~
My defense

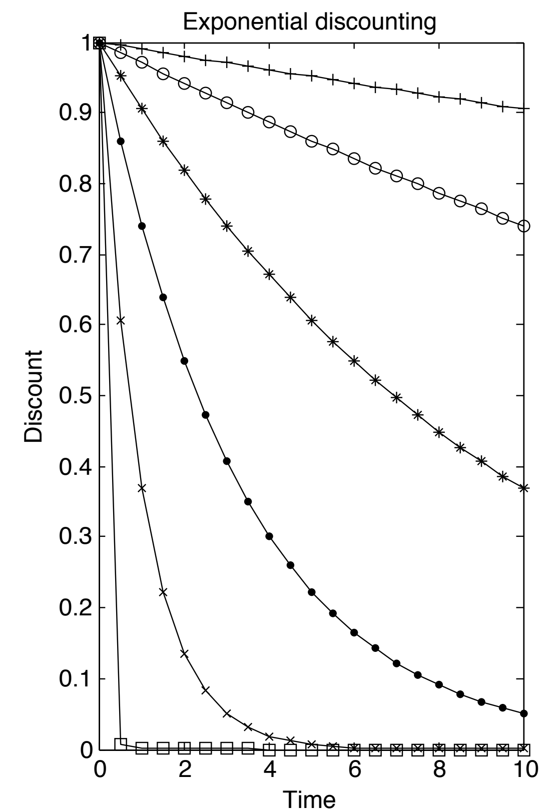


3 Discounting human reward

Reinforcement learning objective is to maximize "long-term" expected reward:

$$\sum_{t=0}^{\infty} E_{\pi}[\gamma^{t} R(s_t, a_t)]$$

discount factor



Discount at t = 0, 1, 2, ………., ∞

γ = 0: 1, 0, 0, ………., 0

γ = 0.5: 1, 0.5, 0.25, …, 0

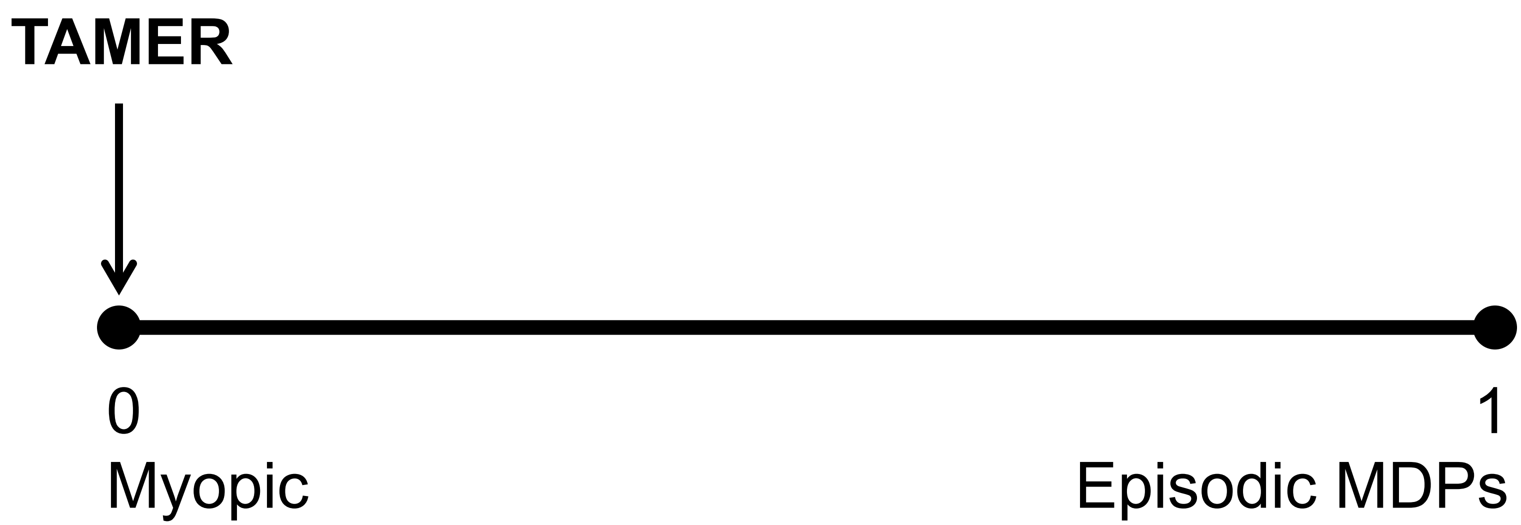
γ = 1: 1, 1, 1, ………., 1

3 Discounting human reward

Discount factor

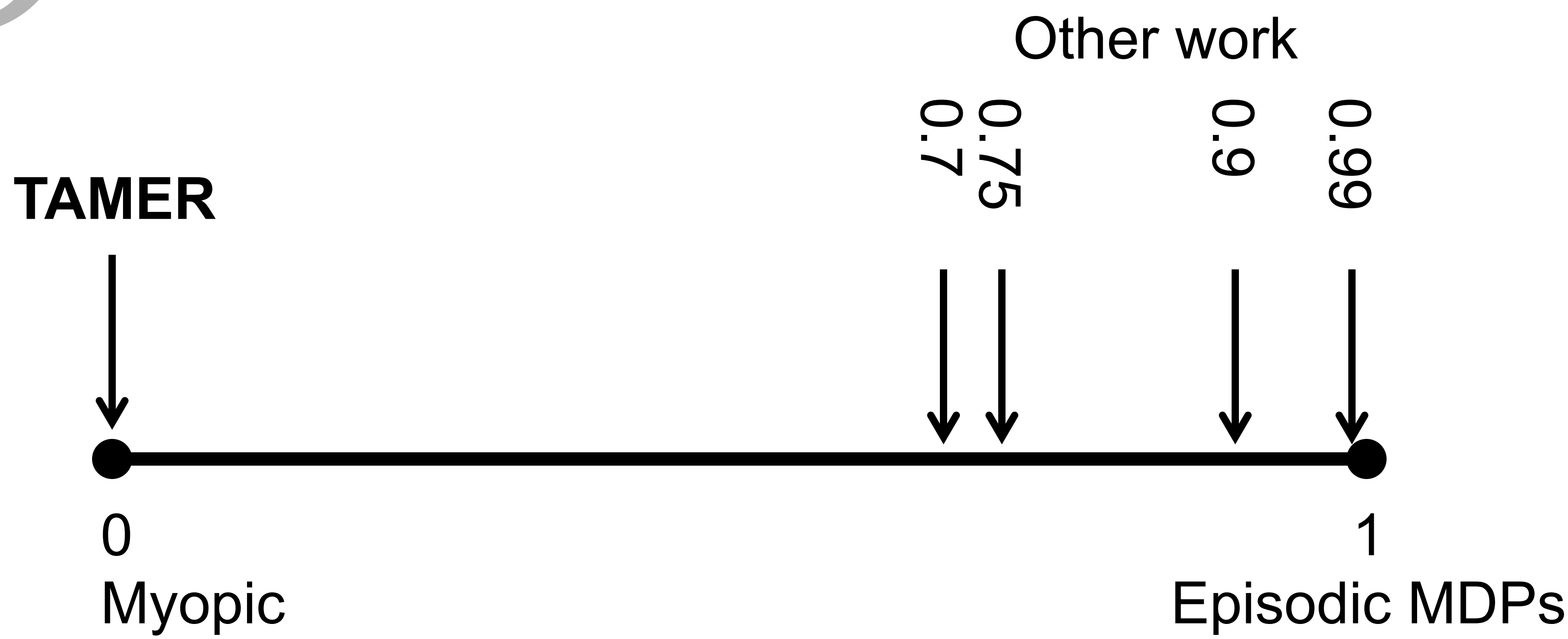
Knox and Stone (2012)

3 Discounting human reward



Discount factor

3 Discounting human reward



Discount factor

Interactive shaping solutions

Earliest publication	Number of tested domains	Effective discount factor	Episodic or continuing tasks tested
Isbell et al. (2000)	1	0.7	Continuing
Thomaz and Breazeal (2006)	1	0.75	Episodic
Knox et al. (2008) - TAMER	6	0	Episodic and continuing
Tenorio-Gonzalez et al. (2010)	2	0.9	Episodic and continuing
Suay and Chernova (2011)	1	0.75	Continuing
Pilarski et al. (2011)	1	0.99	Continuing

Intuition

Positive circuits problem
of goal-based, episodic tasks

Human reward is overwhelmingly positive.

↓

∃ a behavioral circuit with net positive reward.

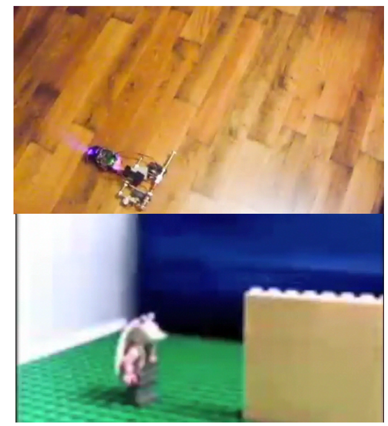
↓

When $\gamma=1$, any (s,a) in circuit is infinitely valued.

↓

Agent never (greedily) reaches the goal.

Example behavioral circuits:



Analysis

Learn with $R \leftarrow \hat{H}$

$$(S, A, T, D, R, \gamma)$$

Analysis

Learn with $R \leftarrow \hat{H}$

$$(S, A, T, D, \hat{H}, \gamma)$$

Analysis

Learn with $R \leftarrow \hat{H}$

$$(S, A, T, D, \hat{H}, \gamma)$$

vary

Analysis

Learn with $R \leftarrow \hat{H}$

$$(S, A, T, D, \hat{H}, \gamma)$$

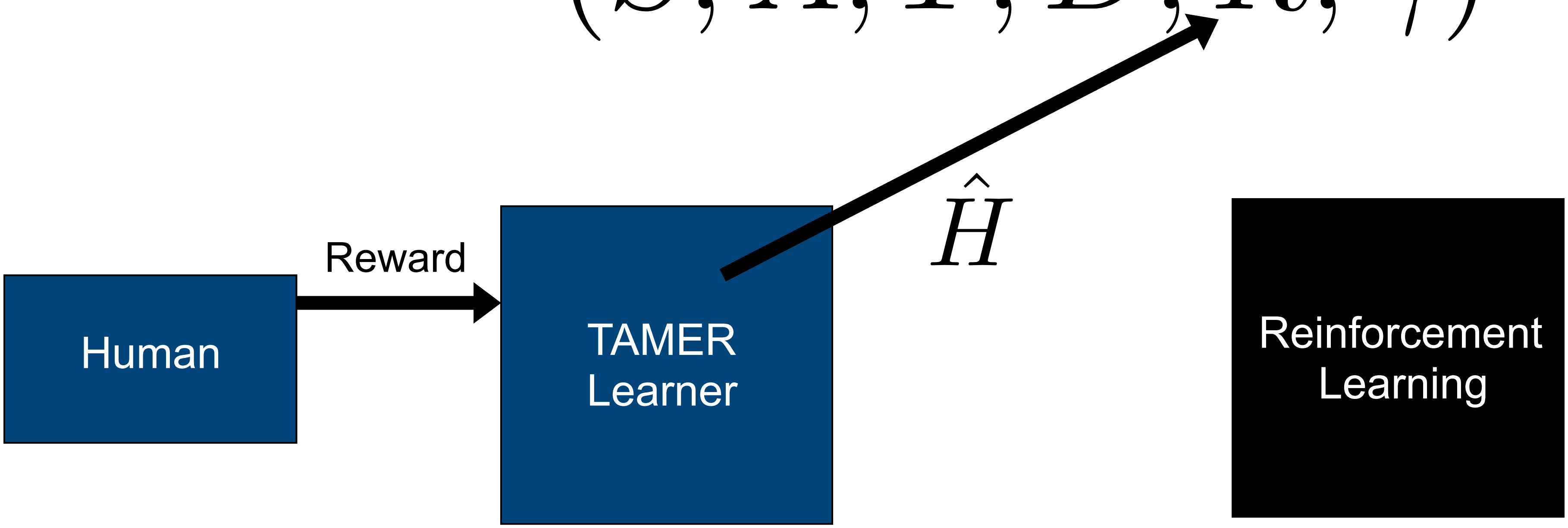
vary

Ask: under what discounting does MDP-optimal behavior translate to best task performance?

Analysis

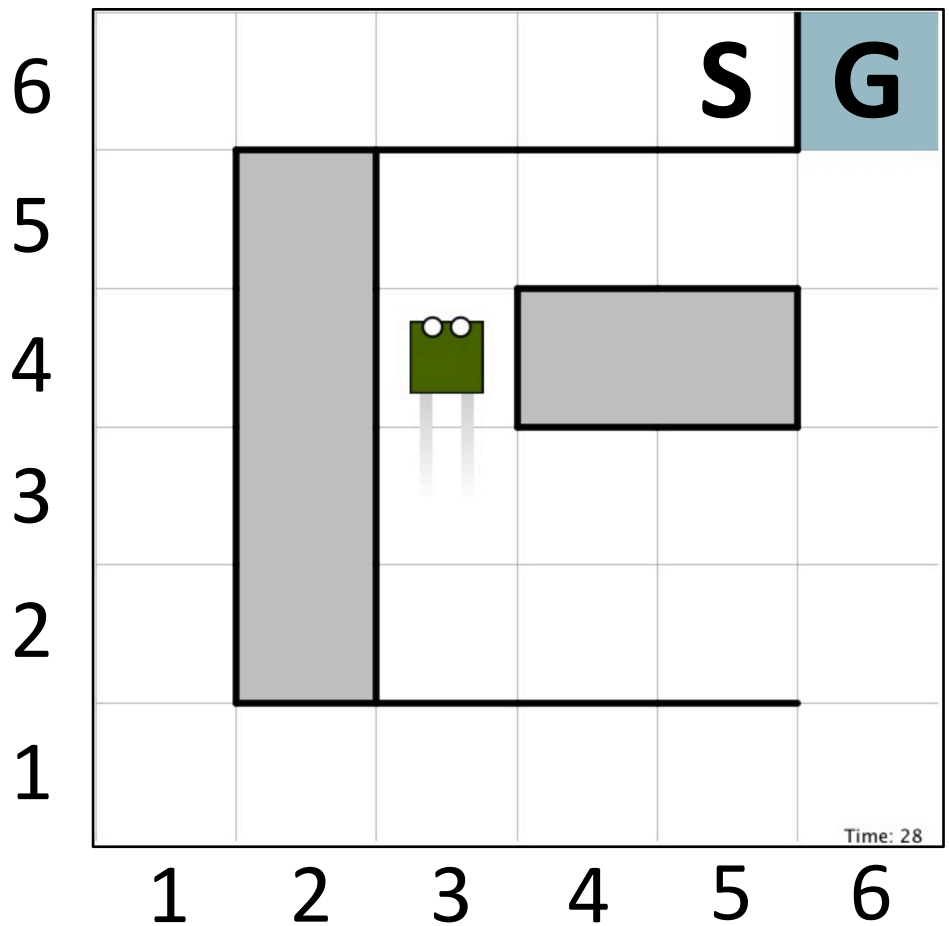
When $\hat{H}$ is trained with actions approximately optimal to MDP

$$(S, A, T, D, R, \gamma)$$



Evidence

When $\hat{H}$ is trained with actions approximately optimal to MDP

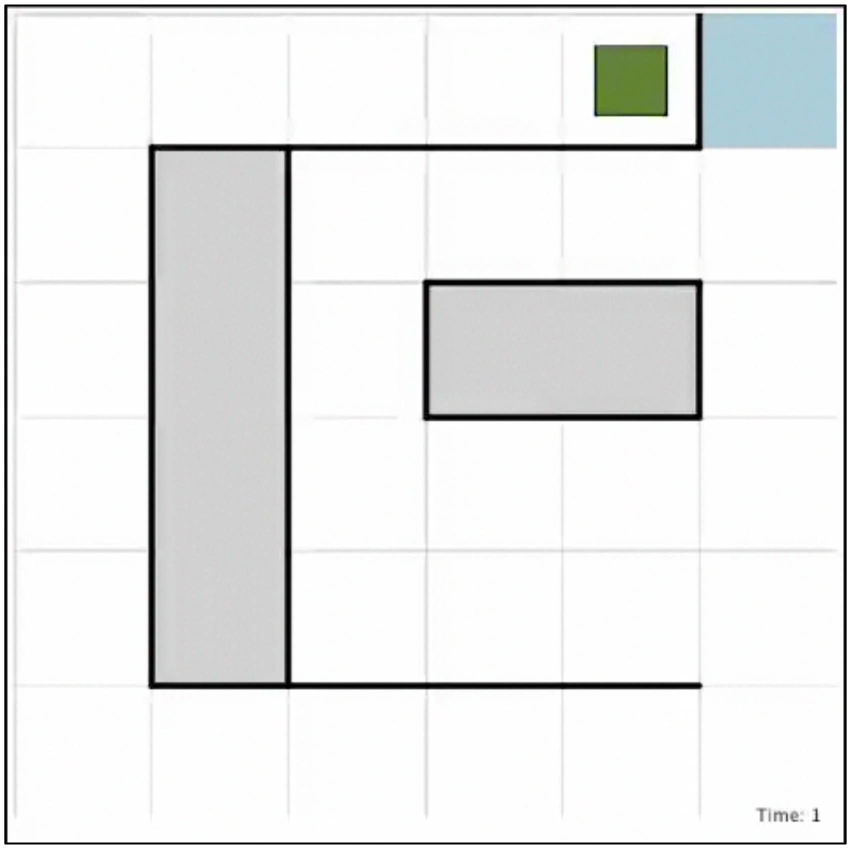


The experiment

- Start and goal states fixed
- 5 episodes or 300 time steps, which comes first
- 7–10 trainers per discount factor

Evidence

When $\hat{H}$ is trained with actions approximately optimal to MDP



Evidence

When $\hat{H}$ is trained with actions approximately optimal to MDP

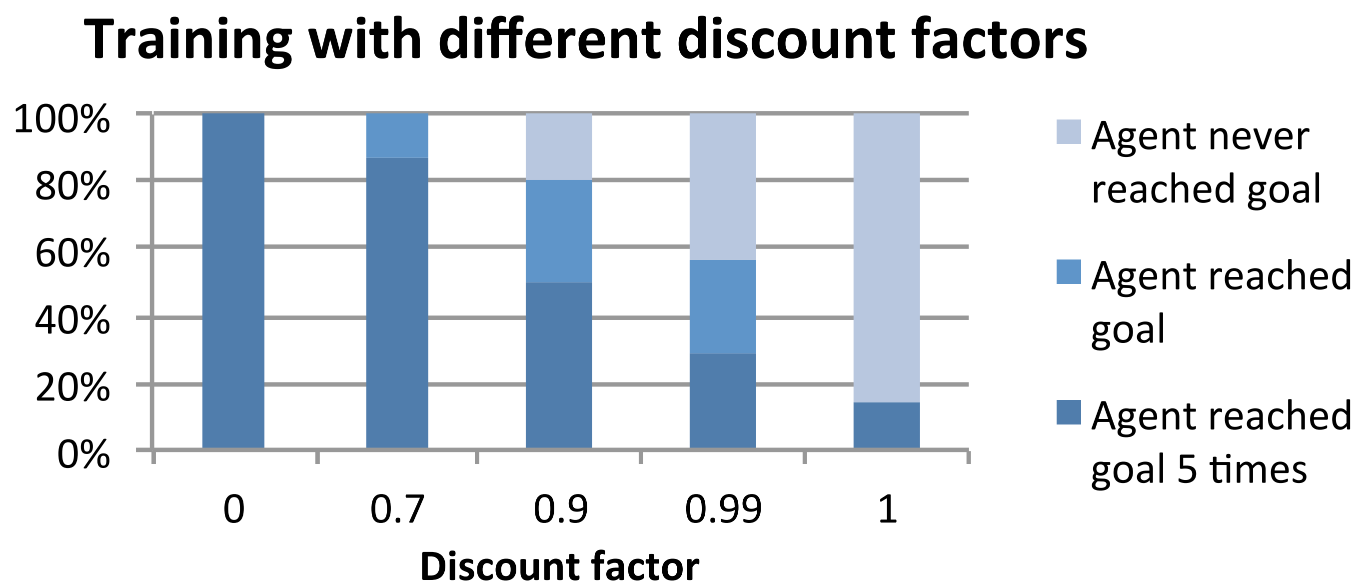
The algorithm:

- One value iteration sweep across states every 20ms
- With 800ms time steps, 40 sweeps per potential change in the reward function

Evidence

When $\hat{H}$ is trained with actions approximately optimal to MDP

Results from Mechanical Turk



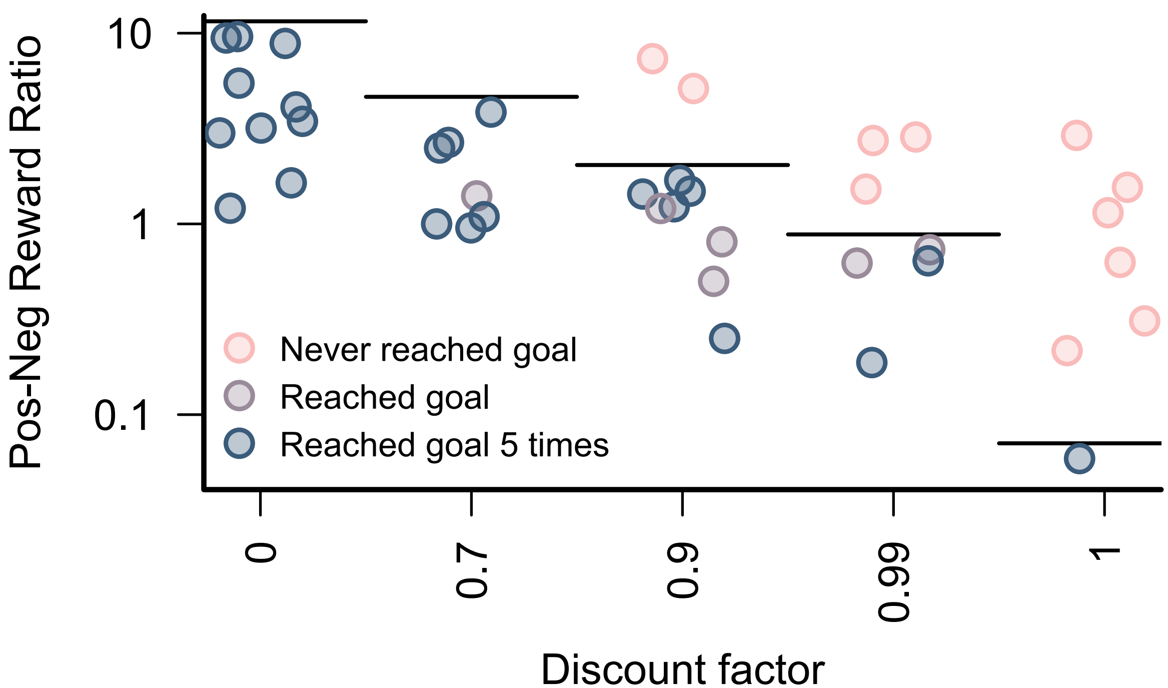
Fisher's Test results (where outcomes are full success or not)

- Comparing 0 and 0.9, $p = 0.0325$
- Comparing 0 and 1, $p = 0.0006$

Evidence

When $\hat{H}$ is trained with actions approximately optimal to MDP

Ratios of positive to negative reward



Observations

1) Reward ratio lowers as γ increases.

2) For a given condition, successful trainers gave more negative reward than unsuccessful trainers.

Evidence

When $\hat{H}$ is trained with actions approximately optimal to MDP

Further observations

3) 66.7% of subjects gave more cumulative positive reward than negative.

4) 83.3% created positive circuits

Evidence

When $\hat{H}$ is trained with actions approximately optimal to MDP

Further observations

3) 66.7% of subjects gave more cumulative positive reward than negative.

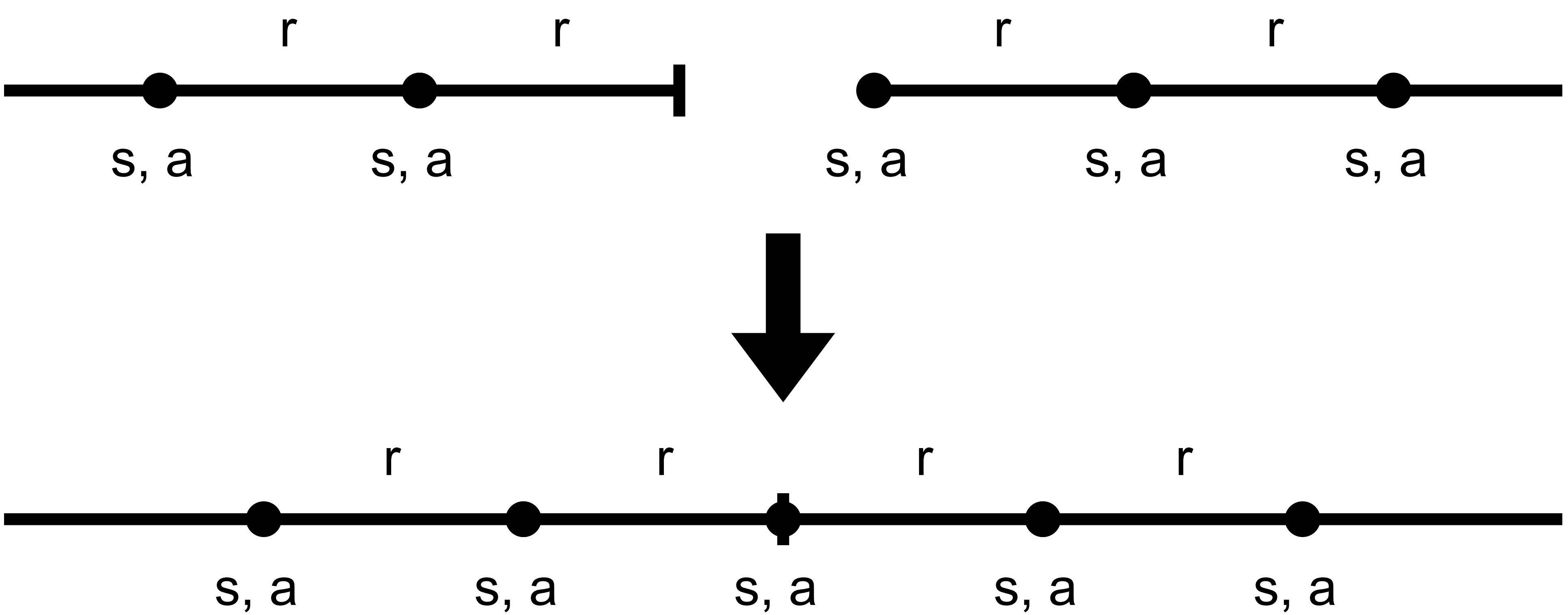
4) 83.3% created positive circuits, ***verifying the prevalence of the positive circuits problem.***

Has TAMER prevailed?

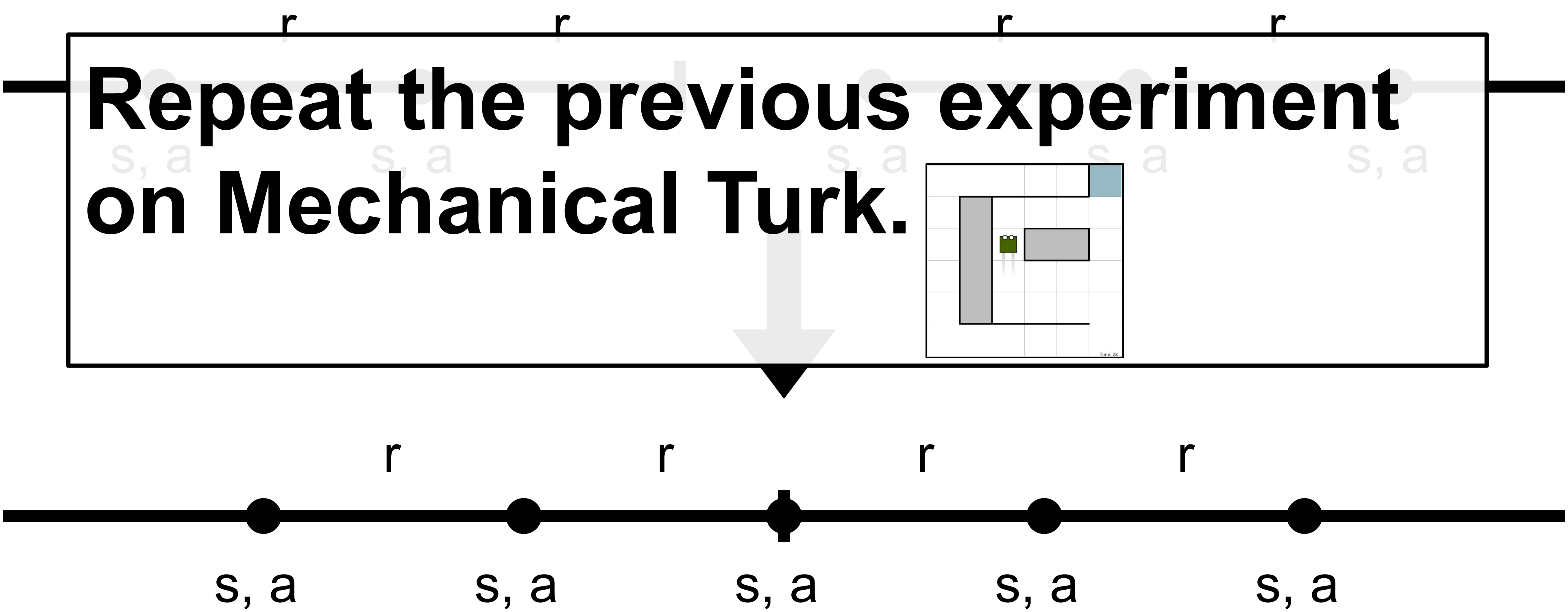
An alternative hypothesis

Continuing tasks do not suffer from the positive circuits problem.

Forcing episodic task to be continuing



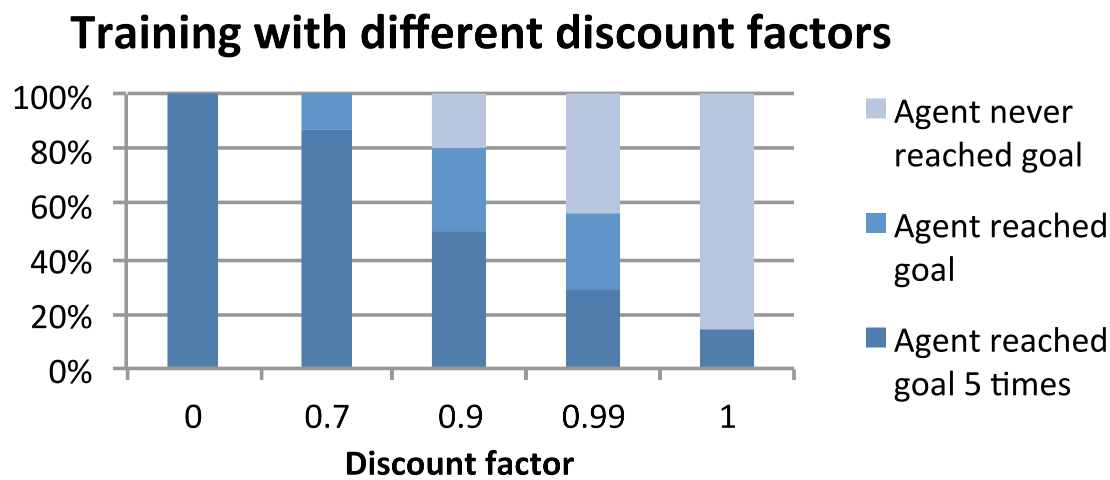
Forcing episodic task to be continuing



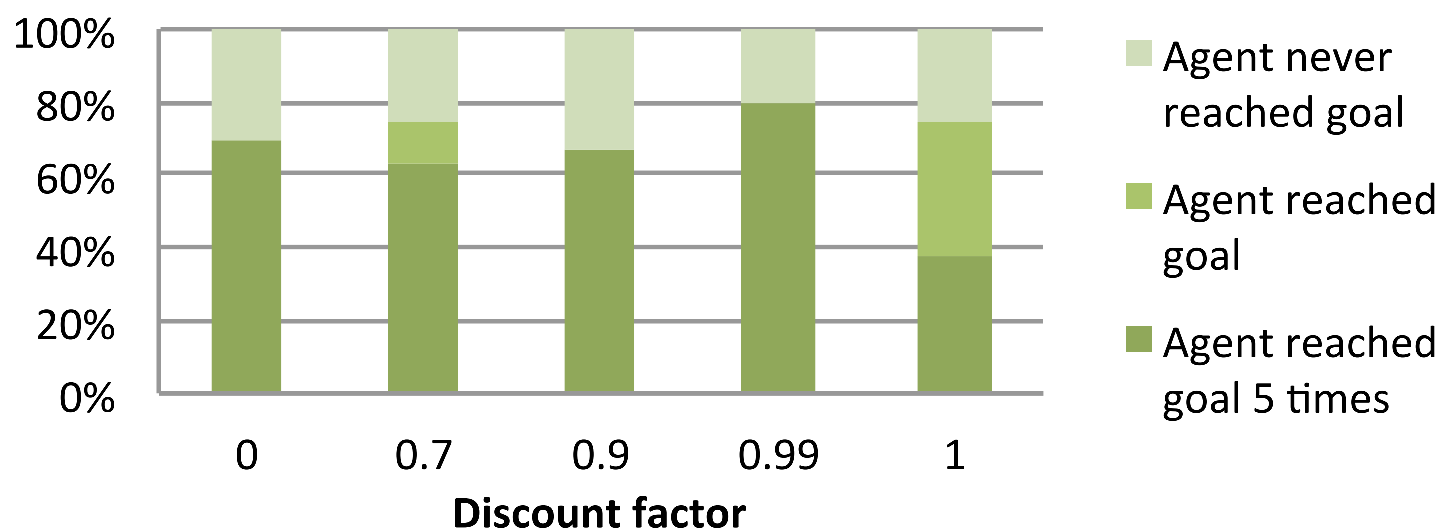
Evidence

Success rates

Episodic (previous):



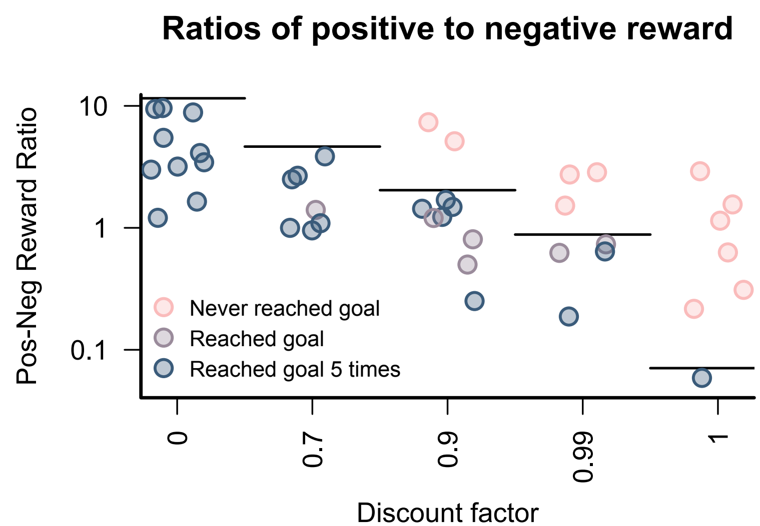
Continuing:

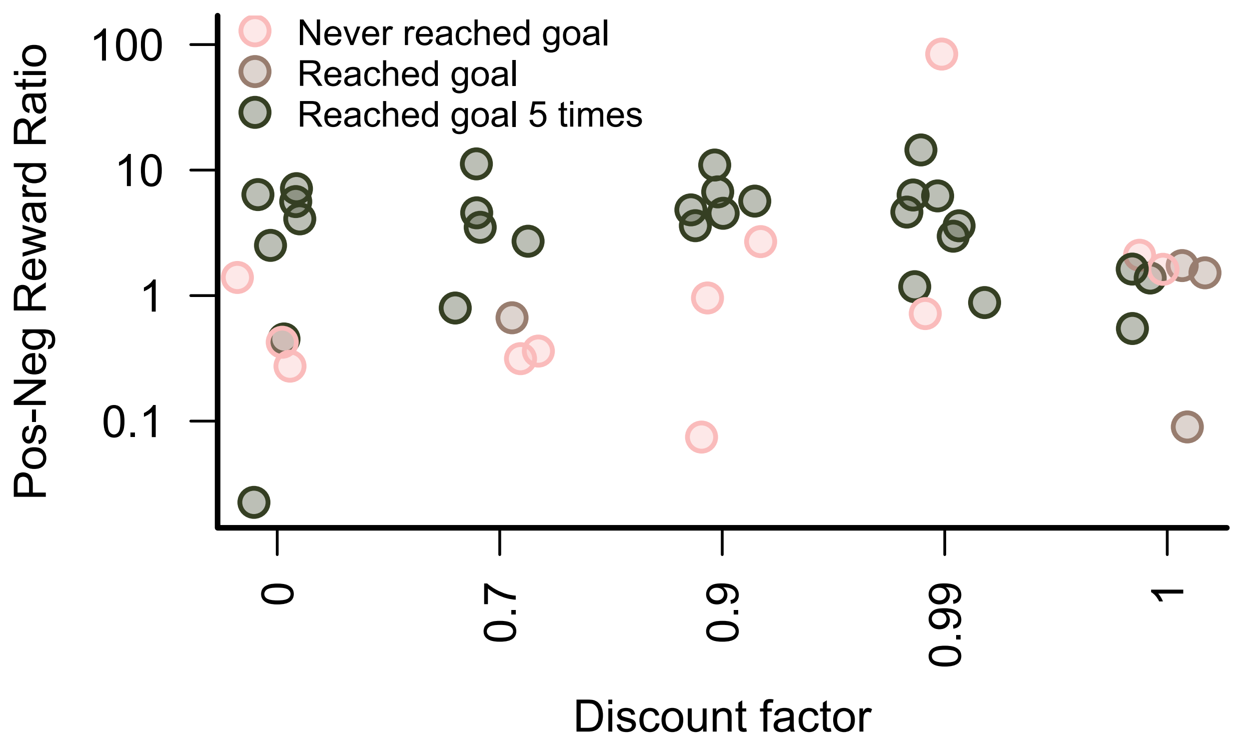


Evidence

Reward positivity

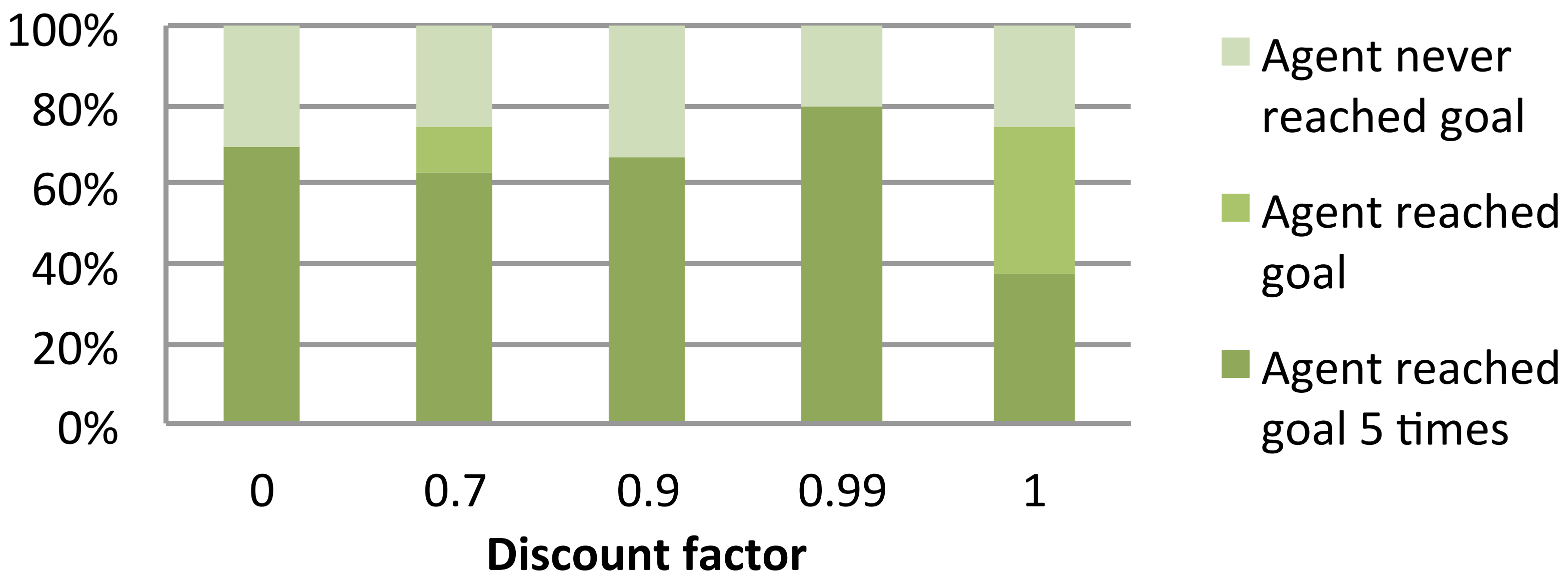
Episodic (previous):





Continuing

Which γ to use then?



Beyond success on this simple, straightforward task, are there other ways to differentiate between γs?

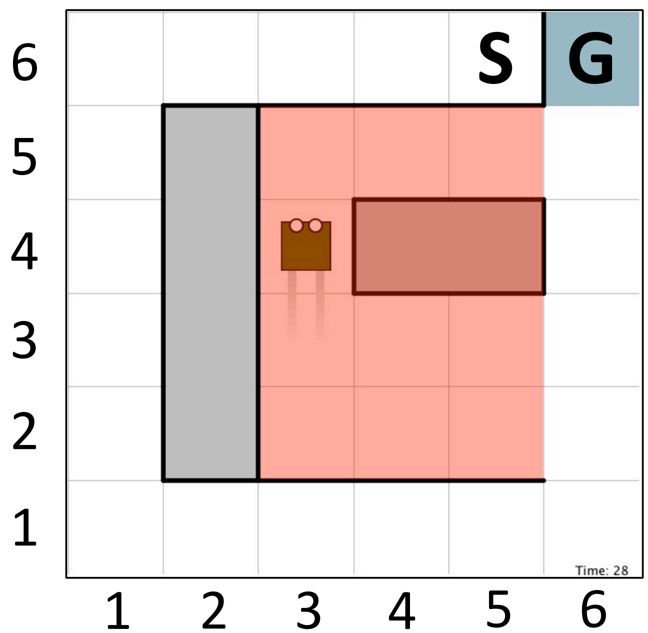
Advantage of low discounting

In theory, task can be communicated (not just policy).

Does it occur in practice?

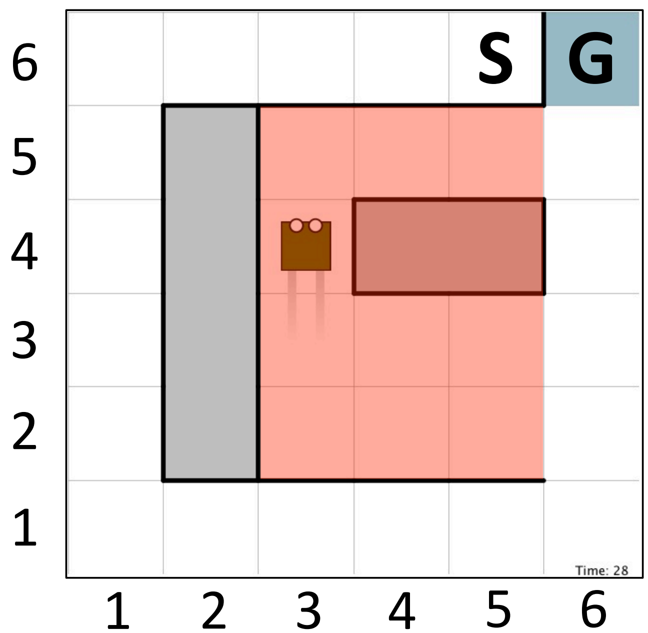
Advantage of low discounting

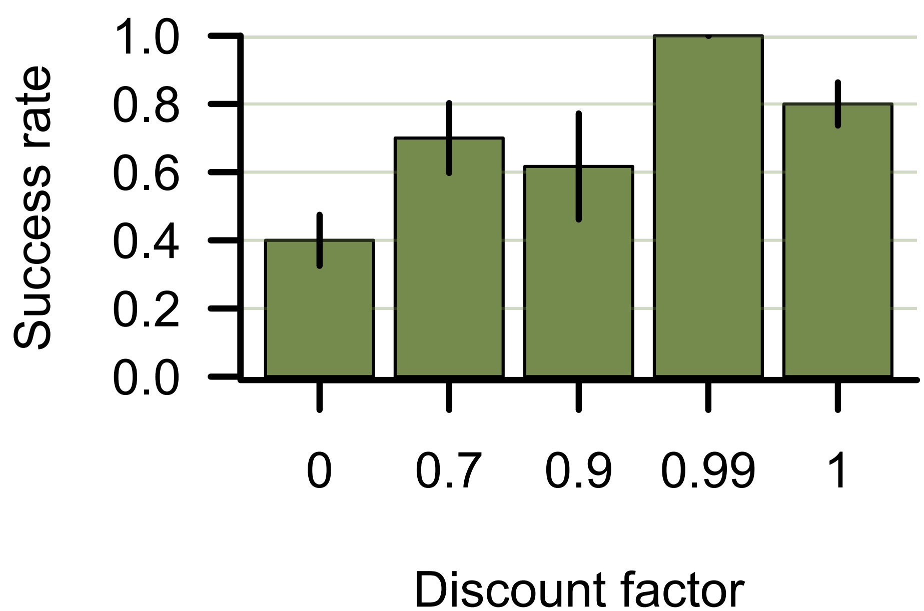
Test 1: Success rate of successfully trained agents **from states off the optimal path**



Advantage of low discounting

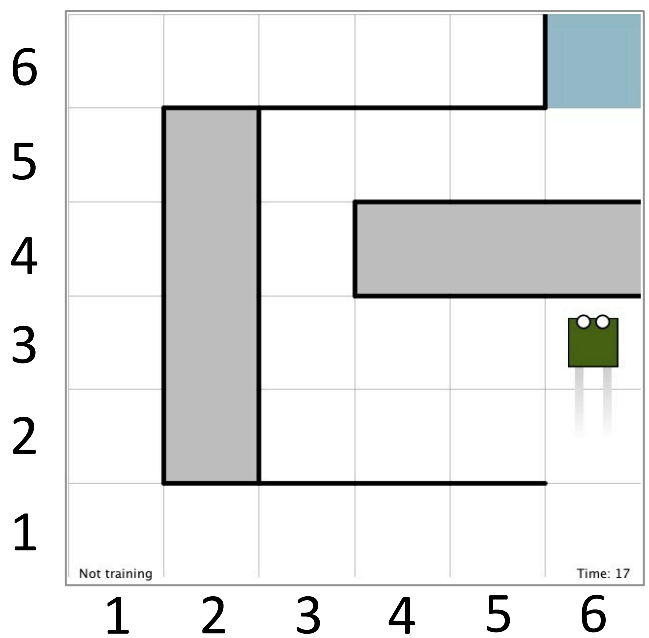
Test 1: Success rate of successfully trained agents **from states off the optimal path**





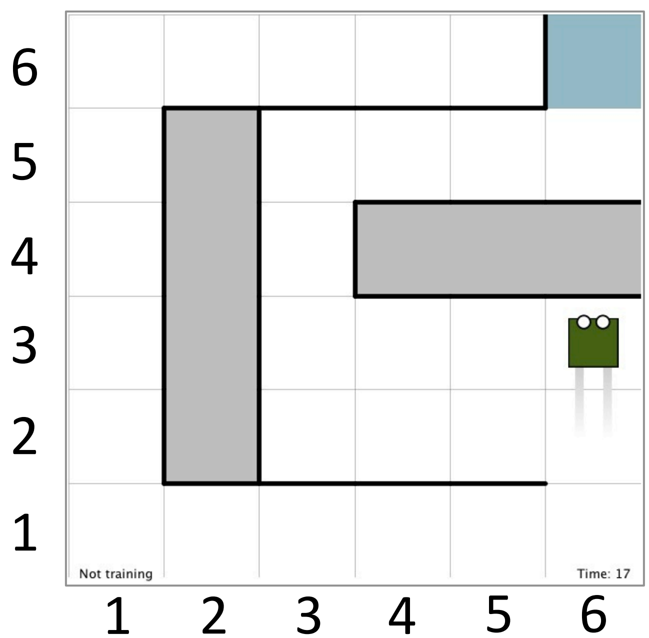
Advantage of low discounting

Test 2: Success rate of successfully trained agents **when optimal path is blocked**



Advantage of low discounting

Test 2: Success rate of successfully trained agents **when optimal path is blocked**



Final trained agents:

4 of 8 $\gamma = 0.99$ agents reach the goal.

No other agents did.

Advantage of low discounting

To some extent, task was communicated (not just policy).

Advantage of low discounting

To some extent, task was communicated (not just policy).

Suggests that interactively shaped agents could learn better policies than those known to the trainer.

Disadvantage of low discounting

In complex tasks with a changing reward function, acting policy is farther from optimal.

In preliminary work, RL algorithms that sample by experience (unlike value iteration) have difficulty learning the simple grid world task.

- Reward-rich world encourages repetition of initial behavior.
- Positive circuits problem only partially solved.

Recommendations for learning from human reward

Therefore,

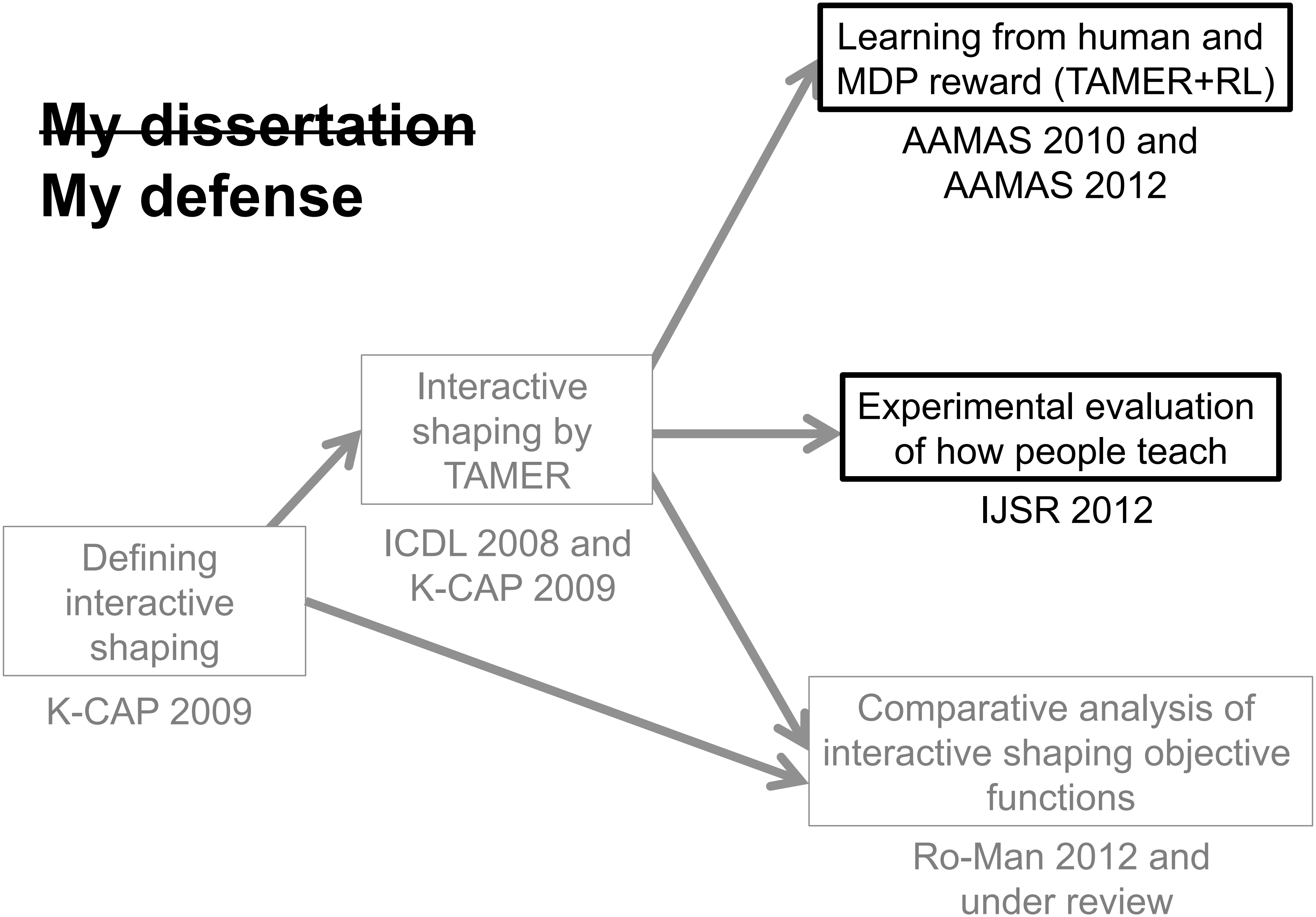
TAMER appears to remain the best current approach,

but algorithms using low discounting in continuing tasks are more promising directions for future work.

Contributions of investigation into discounting

1. Linking human reward positivity to positive circuits, empirically establishing pos. circuits' prevalence, and giving resultant algorithmic guidance
2. Relating γ, human reward positivity, episodicity, and task performance in goal-based tasks
3. The first empirical differentiation of algorithms for learning from human reward
4. First success with low discounting ($\gamma = 0.99$ w/ 0.8 s time steps)

~~My dissertation~~
My defense



4 Learning from human and MDP reward (TAMER+RL)

- Human reward: teaches quickly but imperfectly
- MDP Reward (R): slower learning but specifies optimal behavior

- How to use the two signals together?
  - We test 8 combination techniques.

AAMAS 2010 and AAMAS 2012

4 TAMER+RL **Conclusions**

Human and MDP reward can be combined to improve upon learning from either alone.

Manipulating action selection – highest, most consistent gains and robust to changes in weights

Mixing human and MDP reward in a single value function – sometimes helps, brittle to weight values

5 Experimental evaluation of how people teach

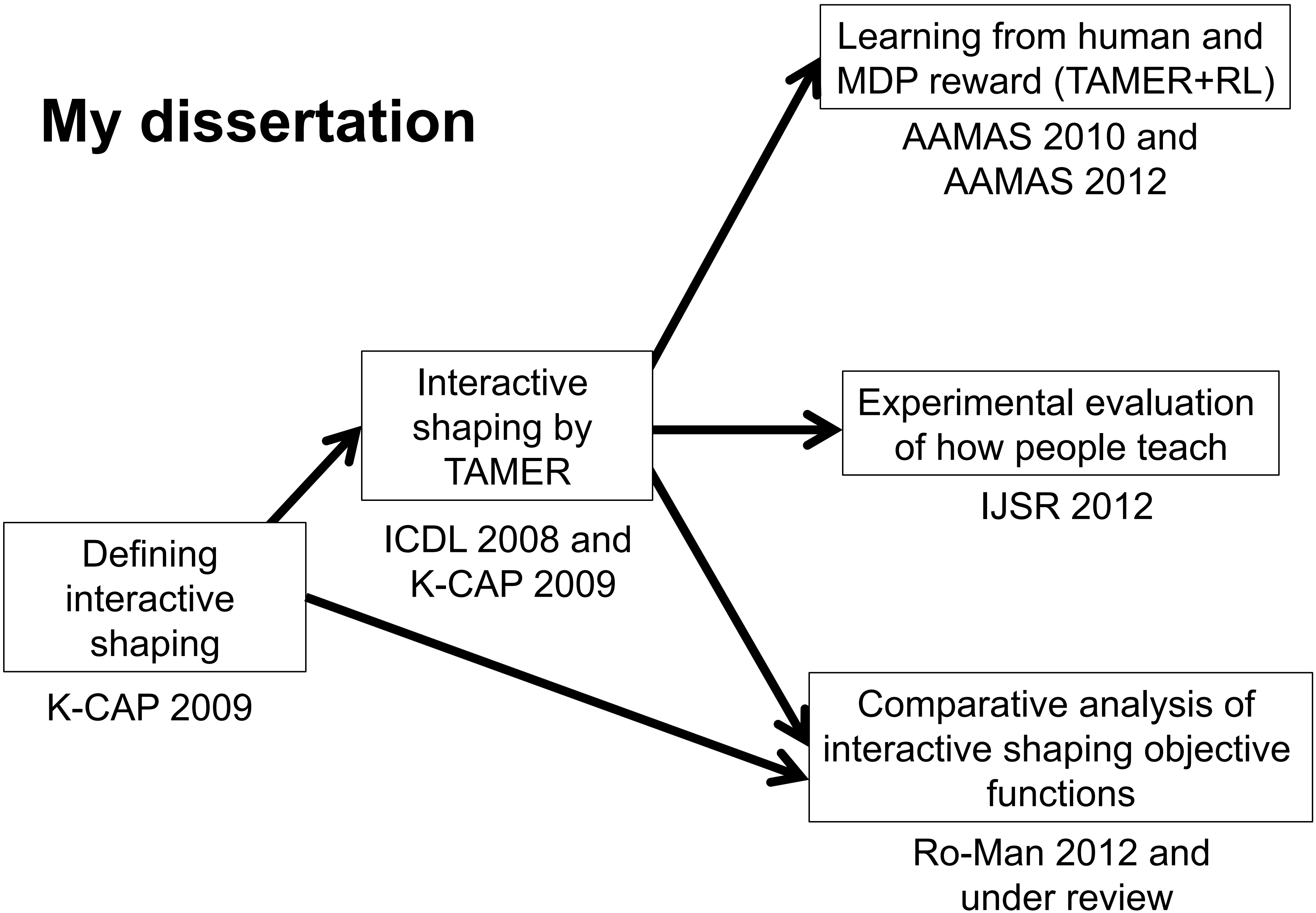
Two well-controlled, relatively large experiments with TAMER agents investigate how the trainer's feedback is impacted

1. by the trainer's self-perceived role and
2. by agent misbehavior.

Early examples of using computational learning agents as highly specifiable social entities in experiments on human behavior.

Knox, Glass, Love, Maddox, and Stone, IJSR 2012

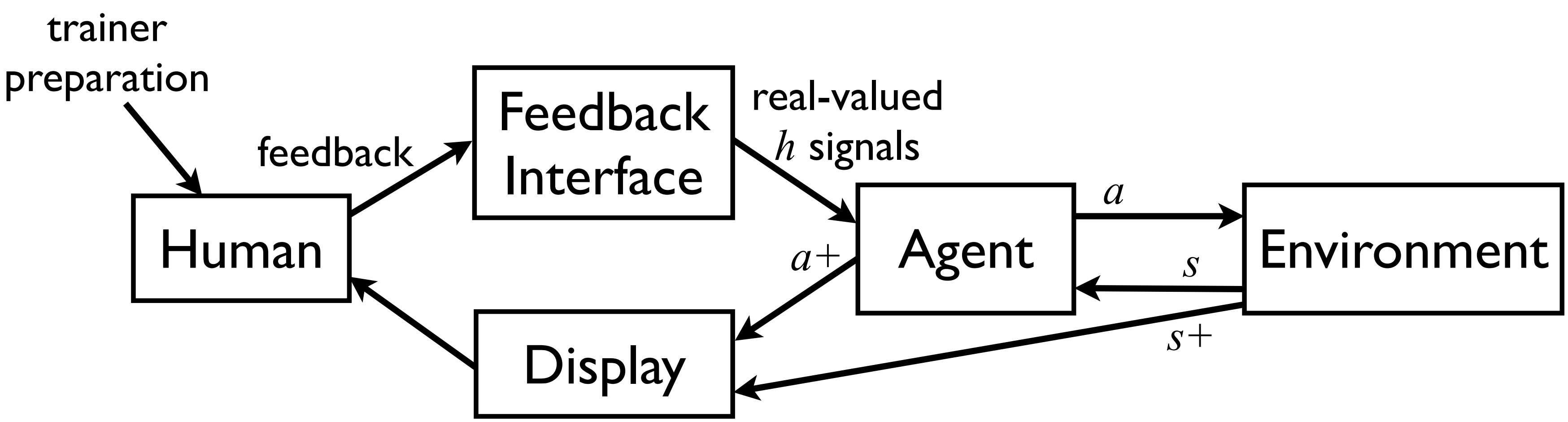
My dissertation



Going forward

Extending this work on pure interactive shaping:

- Trainer preparation
- Transparency
- Interfaces for giving reward
- Mappings from user input to reward values



Going forward

Extending this work on pure interactive shaping:

- Personalization of the learning algorithm
- Biasing towards certain human models
- Non-Markovian models of human reward
- Modeling reward with dimensionality reduction
- Scaling RL algorithms for high γs
- Implement in application domains

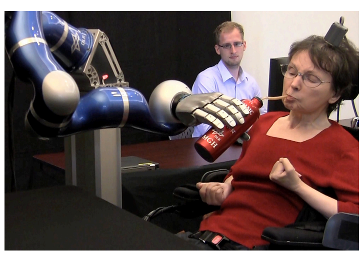
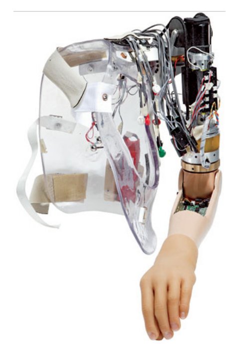
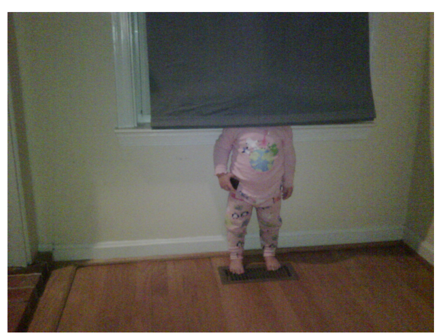


Image is courtesy of ABB



Going forward

Extending *beyond* pure interactive shaping:

- Unintended rewards  
- Revisiting TAMER+RL
- Integrate interactive shaping with other natural teaching methods
- One trainer, multiple agents
- Multiple trainers, one agent
- Hidden state 

Going forward

Research on natural training *makes humans useful to agents.*

- Increase people's control and understanding of agents.
- Increase agents' usability.

The path of AI progress will be determined by what information algorithms can effectively use.

Learning from human reward is about understanding what people want.

Creates a human-centric AI