WASHINGTON STATE UNIVERSITY

World Class. Face to Face.

Learning from human reward

Bei Peng

School of EECS

What is human reward?

Communications of

- approval or disapproval,
- judgment of good or bad behavior or outcomes,
- intention of reward or punishment

can be intuitively mapped to a real-valued signal

Human reward applications

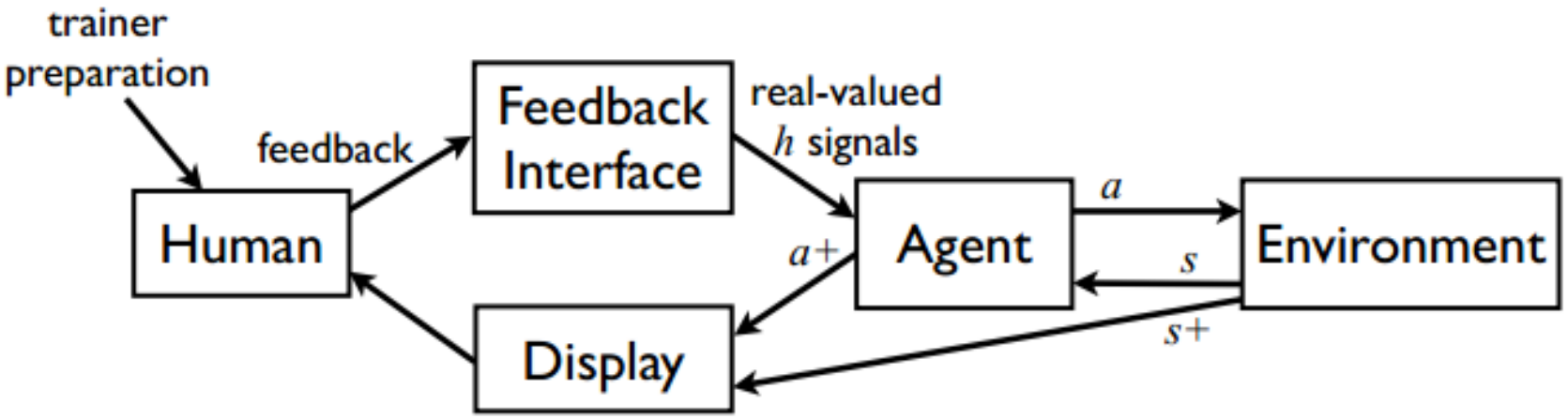
Teaching with human reward

Benefits:

- For undefined tasks, end users can specify correct behaviors

- For defined tasks, human task knowledge can be transferred to ail learning process

Information flow in learning from human reward



Learning from human reward

Human trainer:

- observes agent's behavior
- delivers evaluative reward signals at any time
- attempts to maximize task performance

Agent:

- receives state description
- chooses an action at each time step
- learns from human reward to maximize its performance

Learning from human reward

From the agent's perspective:

- Objective function?
- Correct usage of human reward?
- map human reward to some objective function
- optimize with respect to the objective

Other teaching modalities

- Learning from advice/instruction
  - providing a desire subset of actions when a certain condition is true
  - inaccessible to non-technical users
- Learning from demonstration
  - a special case of instruction
  - demonstrate a task via remote control or his own body

Learning from reward vs. learning from demonstration

LFD	LFR
Requires the human to be an expert or good behavior	Criticism requires less expertise than action
No means for the trainer to see what the agent has learned (problematic behavior)	Agents always express their learned behaviors so that the trainer can adjust feedback
Similar outcome only occurs through randomness or bias	Agent may learn a policy outperform the trainer's intended policy

How to solve the problem of learning from human reward

- Trainer has long-term impact in mind
- Trainer gives reward with small delay

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

- addresses delays through credit assignment
- learns a model of human reward $\widehat{R}_H$
- chooses the action that is predicted to directly elicit the maximum reward $\text{argmax}_a \widehat{R_H}(s, a)$
- different discounting of future human reward

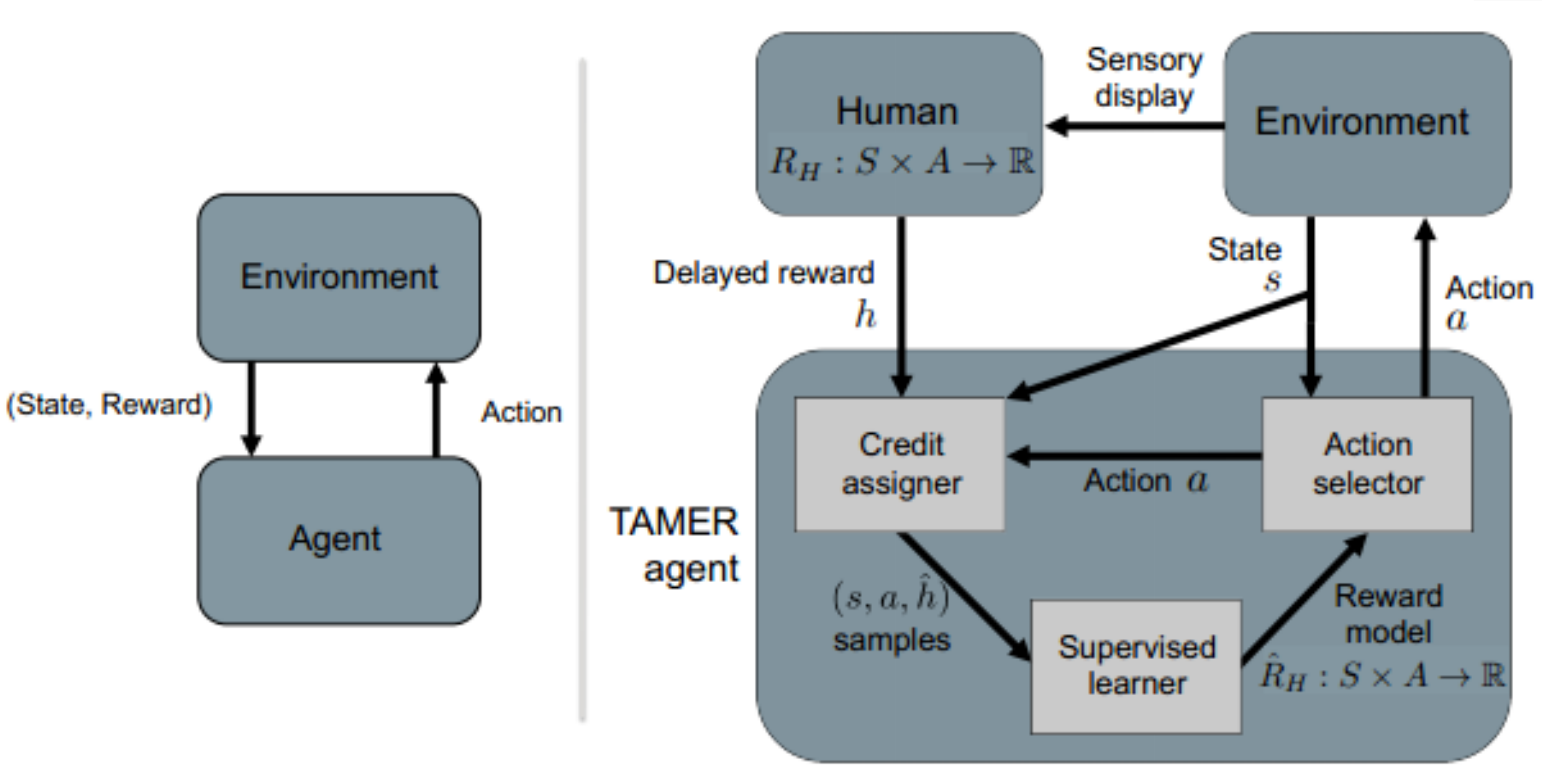
Teaching an Agent Manually via Evaluative Reinforcement (**TAMER**)

- In RL, agents seek to maximize long-turn return, which is a discounted sum of all future reward

- TAMER agents attempt to directly maximize the reward attributable to the chosen state-action pair

  - trainer's reward signal is a direct label on the quality of recent state-action pairs

Teaching an Agent Manually via Evaluative Reinforcement (TAMER)



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

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

Learning R_H is a supervised learning problem

TAMER's process of learning a policy from human reward reduces an apparent reinforcement learning problem to a supervised learning problem with credit assignment

TAMER algorithm

Algorithm 4 The TAMER algorithm for infrequent actions

Main thread

1: $i \leftarrow -1$
2: **while** *true* **do**
3: $i \leftarrow i + 1$
4: **if** *trainer is engaged* **and** $i \geq 2$ **then**
5: $updateModel(\hat{R}_H,\ (e_{i-2}.s, e_{i-2}.a, \hat{h}))$ // Label second-to-last state-action pair
6: **end if**
7: $\hat{h} \leftarrow 0$
8: $e_i.s \leftarrow getState()$
9: $e_i.a \leftarrow argmax_a \hat{R}_H(e_i.s, a)$
10: $takeAction(e_i.a)$
11: wait for next time step
12: **end while**

Human interface thread

1: **while** *true* **do**
2: **if** *new reward* h *with value* $h.v$ *is received* **then**
3: $\hat{h} \leftarrow \hat{h} + h.v$
4: **end if**
5: **end while**

TAMER algorithm

Algorithm 5 The TAMER algorithm

Main thread

Input: ϵ_p

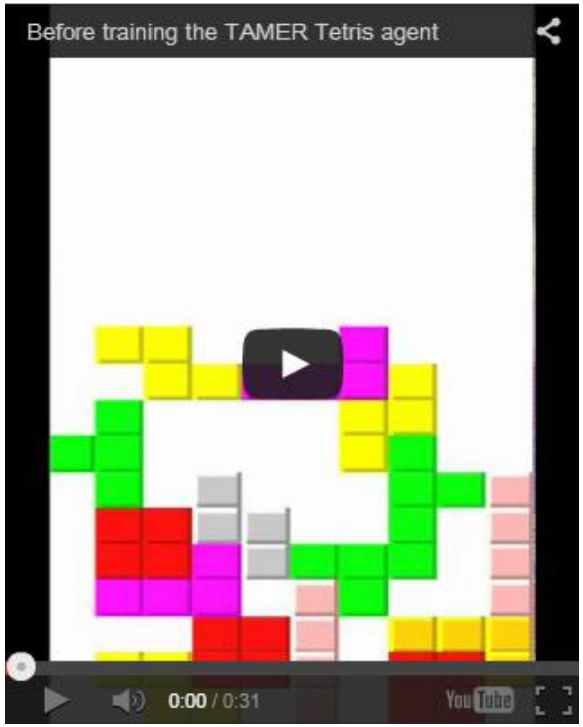
$i \leftarrow -1$
while *true* **do**
  $i \leftarrow i + 1$
  $t_{curr} \leftarrow$ *current time*
  if $i - 1 \geq 0$ **then**
    $e_{i-1}.t_t \leftarrow t_{curr}$
  end if
  $e_i.t_s \leftarrow t_{curr}$
  for all $e \in E_{hist}$ **do**
    if $P(targets(h, e)) < \epsilon_p$ for all $h \in H_{hist}$ such that $h.t > t_{curr}$ **then**
      $e.\hat{h} \leftarrow \sum_{h \in H_{hist}}^{} h.v \int_{e.t_s - h.t}^{e.t_t - h.t} f_{delay}(x)dx$
      $updateModel(\hat{R}_H,\ (e.s, e.a, e.\hat{h}))$ // add completed sample
      $E_{hist} \leftarrow E_{hist} \setminus e$ // remove sample from memory
    end if
  end for
  for all $h \in H_{hist}$ **do**
    if $\int_{e.t_s - h.t}^{0} f_{delay}(x)dx = 1$ for all $e \in E_{hist}$ **then**
      $H_{hist} \leftarrow H_{hist} \setminus h$ // remove fully credited reward from memory
    end if
  end for
  $e_i.s \leftarrow getState()$
  $e_i.a \leftarrow argmax_a \hat{R}_H(e_i.s,\ a)$
  $E_{hist} \leftarrow E_{hist} \cup \{e_i\}$ // record new experience
  $takeAction(e_i.a)$
  wait for next time step
end while

Human interface thread

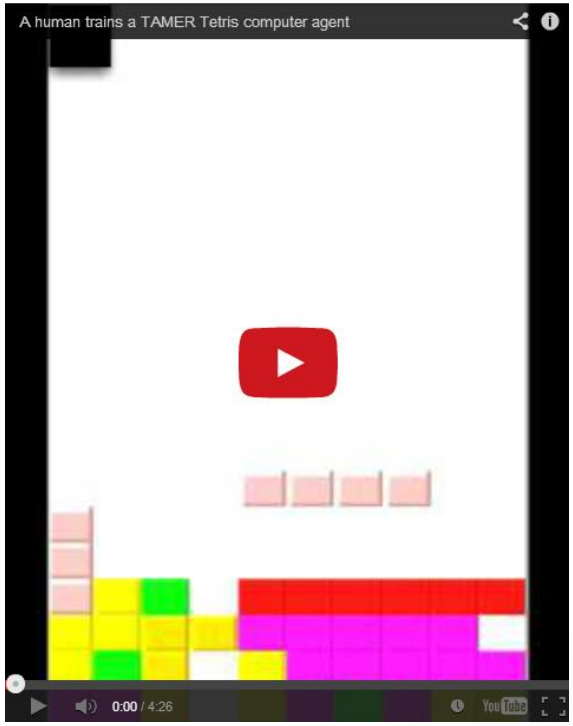
while *true* **do**
  if *new reward h with value h.v is received* **then**
    $h.t \leftarrow$ *current time*
    $H_{hist} \leftarrow H_{hist} \cup \{h\}$ // add new reward signal
  end if
end while

TAMER: Tetris

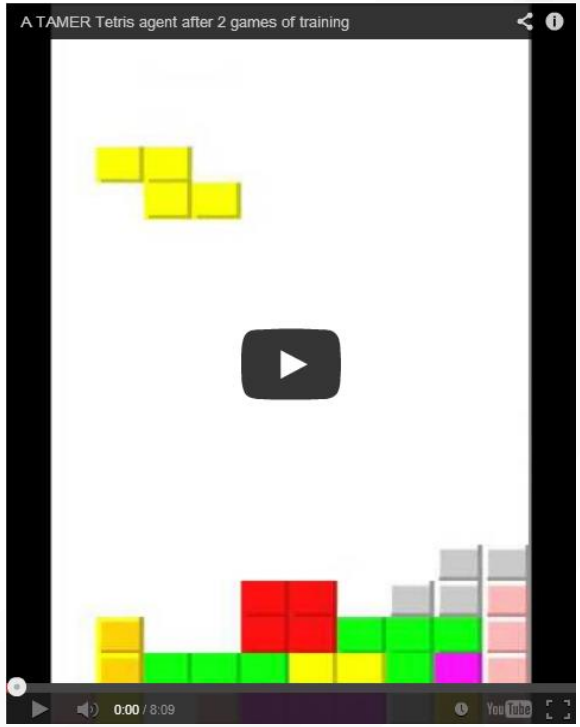
Before training:



During training:



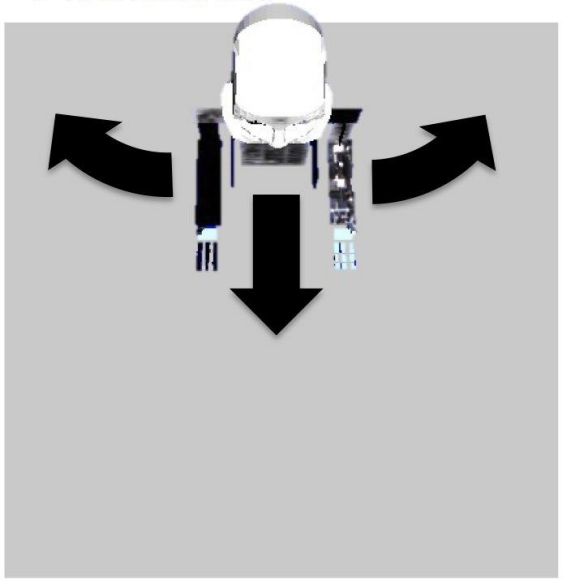
After training:



http://www.bradknox.net/human-reward/tamer/tamer-in-action/tetris/

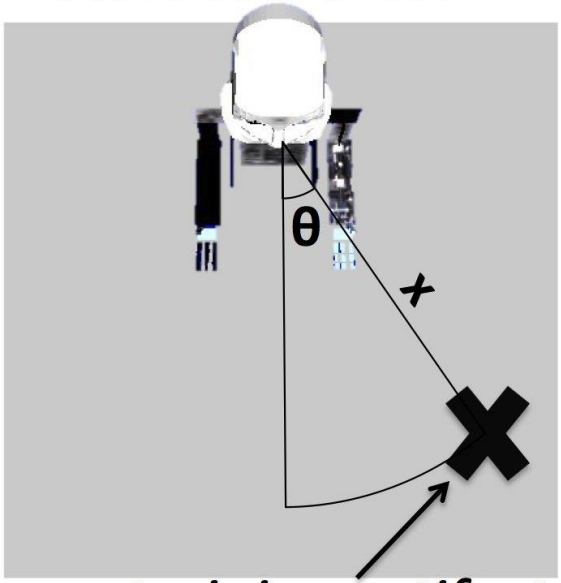
TAMER: interactive robotics

4 actions:



and “stay”

2 state features:

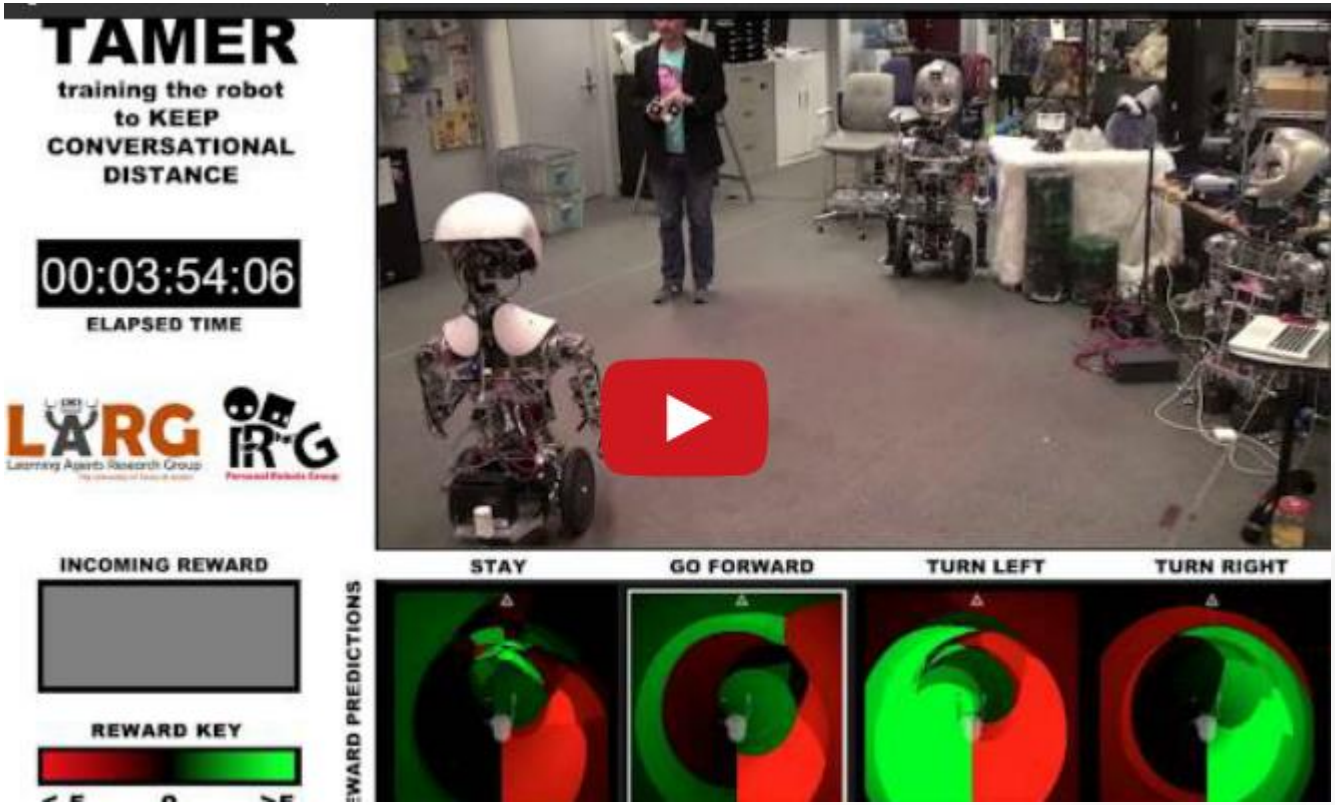


training artifact

Reward interface:

Knox, Stone, and Breazeal, 2012

TAMER: interactive robotics



http://www.bradknox.net/human-reward/tamer/tamer-in-action/interactive-robot-navigation/

TAMER results

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

TAMER agents learn good behavior faster
and
learners using MDP reward meet or surpass TAMER agents’ performances when given enough learning samples

The MDP reward signal

Sparse and delayed

- discriminating reward is rarely received
- wait until the end of game to determine the quality of the state-action pair

MDP reward only *indirectly* evaluates behavior

The human reward signal

Not sparse and delayed

- each reward fully distinguishes between approved and disapproved behavior
- can be delivered with trivial delay
- though flawed, can efficiently learn good behavior

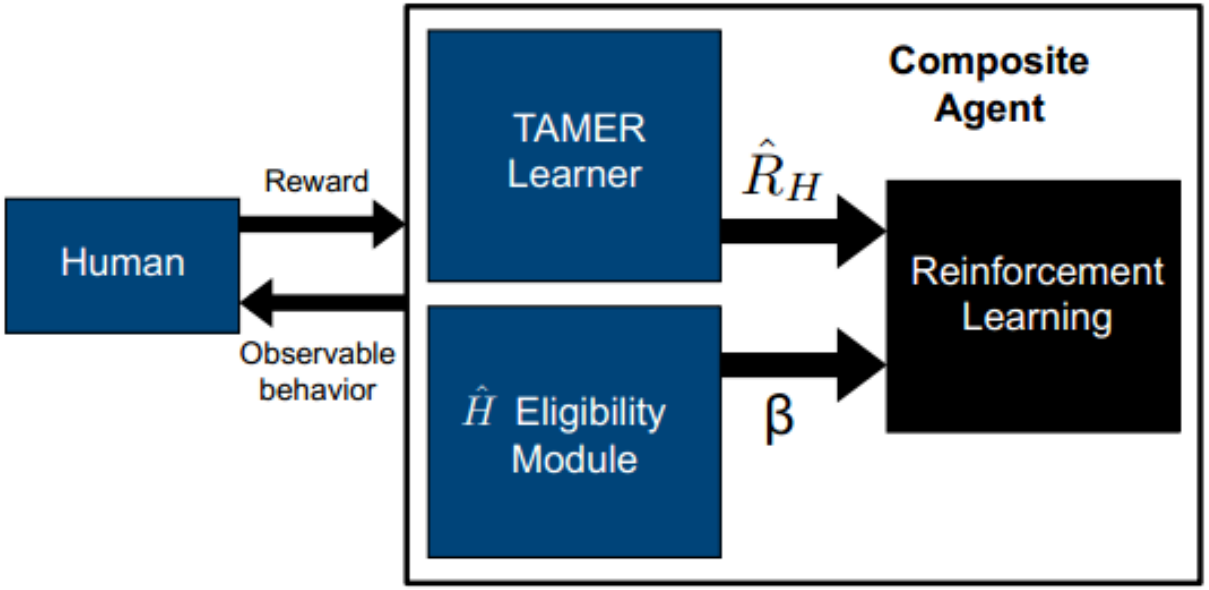
Human reward *directly* evaluates behavior

Sequential TAMER + RL

TAMER: fast learning

RL: learn better policy (long-term)

Simultaneous TAMER + RL



A trainer can step-in as desired to change the course of learning at any time

Future work

- Trainer preparation
- Interfaces for giving reward (push-buttons ?)
- Mappings from user input to reward values (+1 and -c)
- Personalization of the learning algorithm
- Implement in application domains

Future work

Learning from human reward is about understanding *what people want* and how to *provide these outcomes* that people desire

AI is often dominated by vision of autonomy

Creates human-centered AI

- makes humans useful to agents
- understand and control the behavior of agents

Thanks

WASHINGTON STATE UNIVERSITY

World Class. Face to Face.