**Shaping**
- Human trainer transfers task knowledge to an agent through signals of positive and negative reinforcement (shaping the agent).
- How should the agent use human reinforcement signals to learn the task, maximizing performance?

**Why shape?**
- Potential benefits over "autonomous" learning agents include:
  1. no coded evaluation function needed
  2. decreases sample size for learning a "good" policy
  3. allows users to teach agents policies which they prefer
  4. can learn in domains that are intractable for evaluation function methods

**The shaped agent's perspective**
Each time step, agent:
- receives state description
- might receive a scalar human reinforcement signal
- chooses an action
- does not receive an MDP reward signal

**Previous work on shaping agents**
- Clicker training for entertainment agents
  - RL with reward = environmental (MDP) reward + human reinforcement
- Sophie’s World
  - RL with reward = environmental (MDP) reward + human reinforcement
- Social software agent Cobot in LambdaMoo
  - RL with reward = human reinforcement

**MDP reward vs. human reinforcement**

**Human reinforcement**
- Trainer has long-term impact in mind
- Reinforcement is within a small temporal window of the targeted behavior
- Credit assignment problem is largely removed

**The TAMER Framework**
Training an Agent Manually via Evaluative Reinforcement (TAMER)
- Human trainer’s reinforcement reflects his understanding of behavior’s long-term consequences
- Therefore, no temporal difference learning necessary
- Model human’s reinforcement function, $H: S \times A \rightarrow R$
- Exploit learned $H$ to choose actions

**Tetris Results**
Results of various Tetris agents:

**Tetris and TAMER**
- Drop blocks to make solid horizontal lines, which then disappear
- (state space) $> 2^{35}$
- Challenging (NP hard) but infrequent actions
- 46 features extracted from state and action
- Linear model over features
- Gradient descent updates from error $h \rightarrow \hat{h}$
- Greedy action selection

**Mountain Car Results**

**Why shape?**
- Drop blocks to make solid horizontal lines, which then disappear
- Challenging (NP hard) but infrequent actions
- 46 features extracted from state and action
- Linear model over features
- Gradient descent updates from error $h \rightarrow \hat{h}$
- Greedy action selection

**The mean number of lines cleared per game by TAMER agents:**

**References**

**Publications**

**MDP Reward**
- Key problem: credit assignment from sparse rewards
- How did you win?
  - I won!
- But why did I win?

**Tetris Results**

**Mean Cumulative Reward in Mountain Car**

**Mean Cumulative Reward in Mountain Car**