Making Friends on the Fly: Advances in Ad Hoc Teamwork

Samuel Barrett

University of Texas at Austin
sbarrett@cs.utexas.edu

Thesis Defense
October 29, 2014

Acknowledgments

Advisor:

Peter Stone

Collaborators:

Noa Agmon

Noam Hazon

Sarit Kraus

Avi Rosenfeld

Funding: NDSEG Fellowship, NSF, ONR, FHWA

Example

Credit: www.nimsonline.com/impact-of-earthquakes.html

Credit: NIST

Example


Credit: www.nimsonline.com/impact-of-earthquakes.html


Credit: NIST

Example

Credit: www.nimsonline.com/impact-of-earthquakes.html

Credit: NASA

Credit: NIST

Learning Agents Research Group
The University of Texas at Austin

Example

Credit: www.nimsonline.com/impact-of-earthquakes.html

Credit: NASA

Credit: NIST

Example

Credit: www.nimsonline.com/impact-of-earthquakes.html

Credit: NASA

Credit: Nicolas Halftermeyer

Learning Agents Research Group
The University of Texas at Austin

Ad Hoc Teamwork

- Only in control of a single agent
- Unknown teammates
- Shared goals
- No pre-coordination

Examples in humans:

- Pick up soccer
- Accident response

Credit: Soccer Toronto

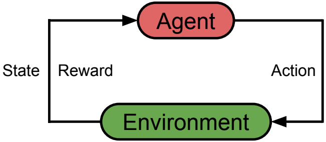
Credit: Shuets Udono

Motivation

- Agents are becoming more common and lasting longer
  - Both robots and software agents
- Pre-coordination may not be possible
- Agents should be robust to various teammates
- Need to adapt quickly!

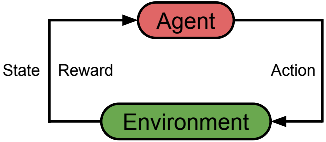
What have people done in the past?

Single agent learning



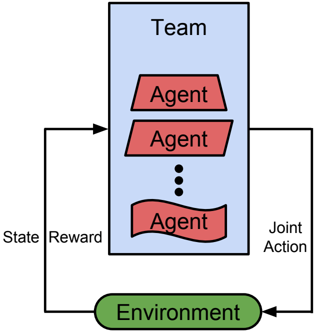
- Existing research shows how a singe agent can effectively learn about its environment
- [Watkins 1989], [Ernst et al. 2005], [Sutton and Barto 1998]

Single agent learning



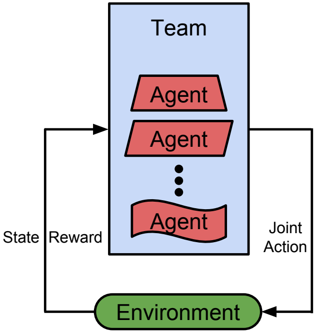
- Existing research shows how a singe agent can effectively learn about its environment
- [Watkins 1989], [Ernst et al. 2005], [Sutton and Barto 1998]
- Assumes that the agent is alone

Multiagent Coordination



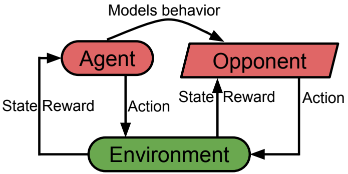
- Existing research provides protocols for coordinating and communicating multiple agents to accomplish their tasks
- [Tambe 1997], [Decker and Lesser 1995], [Lauer and Riedmiller 2000]

Multiagent Coordination



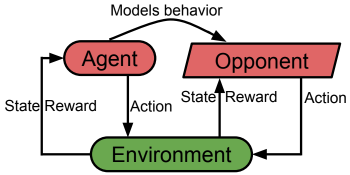
- Existing research provides protocols for coordinating and communicating multiple agents to accomplish their tasks
- [Tambe 1997], [Decker and Lesser 1995], [Lauer and Riedmiller 2000]
- Assumes that all agents share a coordination algorithm

Opponent Modeling



- Learn about opponents through interactions
- [Conitzer and Sandholm 2007], [Korzhyk et al. 2011], [Bard et al. 2013]

Opponent Modeling



- Learn about opponents through interactions
- [Conitzer and Sandholm 2007], [Korzhyk et al. 2011], [Bard et al. 2013]
- Assume the worst case: the other agents are trying to exploit our agent

Ad Hoc Teamwork

Paper	Control Agent	Multiple Teammates	Unknown Teammates	Evaluated in a Complex Domain	Generally Applicable	Adapt Quickly	Automatically Reuse Knowledge
Stone and Kraus (2010)	**Yes**	No	No	No	No	**Yes**	No
Barrett and Stone (2011)	**Yes**	No	No	No	No	**Yes**	No
Brafman and Tennenholtz (1996)	**Yes**	No	No	No	No	No	No
Stone et al. (2010)	**Yes**	No	No	No	No	**Yes**	No
Agmon and Stone (2012)	**Yes**	**Yes**	No	No	No	**Yes**	No
Agmon et al. (2014)	**Yes**	**Yes**	Partially	No	No	**Yes**	No
Chakraborty and Stone (2013)	**Yes**	No	**Yes**	No	No	No	No
Hao et al. (2014)	**Yes**	**Yes**	No	No	No	No	No
Wu et al. (2011)	**Yes**	No	**Yes**	No	Partially	No	No
Albrecht and Ramamoorthy (2013)	**Yes**	**Yes**	Partially	**Yes**	**Yes**	**Yes**	No
Wray and Thompson (2014)	**Yes**	**Yes**	No	**Yes**	No	**Yes**	No
Bowling and McCracken (2005)	**Yes**	**Yes**	Partially	**Yes**	No	No	No
Jones et al. (2006)	**Yes**	**Yes**	No	**Yes**	Partially	**Yes**	No
Su et al. (2014)	**Yes**	**Yes**	No	**Yes**	No	**Yes**	No
Han et al. (2006)	**Yes**	**Yes**	No	**Yes**	No	**Yes**	No
Genter et al. (2013)	**Yes**	**Yes**	No	**Yes**	No	**Yes**	No
Genter and Stone (2014)	**Yes**	**Yes**	No	**Yes**	No	**Yes**	No
Liemhetcharat and Veloso (2014)	No	**Yes**	**Yes**	**Yes**	**Yes**	No	No

LARG
Learning Agents Research Group
The University of Texas at Austin

Ad Hoc Teamwork

Paper	Control Agent	Multiple Teammates	Unknown Teammates	Evaluated in a Complex Domain	Generally Applicable	Adapt Quickly	Automatically Reuse Knowledge
Stone and Kraus (2010)	**Yes**	No	No	No	No	**Yes**	No
Barrett and Stone (2011)	**Yes**	No	No	No	No	**Yes**	No
Brafman and Tennenholtz (1996)	**Yes**	No	No	No	No	No	No
Stone et al. (2010)	**Yes**	No	No	No	No	**Yes**	No
Agmon and Stone (2012)	**Yes**	**Yes**	No	No	No	**Yes**	No
Agmon et al. (2014)	**Yes**	**Yes**	Partially	No	No	**Yes**	No
Chakraborty and Stone (2013)	**Yes**	No	**Yes**	No	No	No	No
Hao et al. (2014)	**Yes**	**Yes**	No	No	No	No	No
Wu et al. (2011)	**Yes**	No	**Yes**	No	Partially	No	No
Albrecht and Ramamoorthy (2013)	**Yes**	**Yes**	Partially	**Yes**	**Yes**	**Yes**	No
Wray and Thompson (2014)	**Yes**	**Yes**	No	**Yes**	No	**Yes**	No
Bowling and McCracken (2005)	**Yes**	**Yes**	Partially	**Yes**	No	No	No
Jones et al. (2006)	**Yes**	**Yes**	No	**Yes**	Partially	**Yes**	No
Su et al. (2014)	**Yes**	**Yes**	No	**Yes**	No	**Yes**	No
Han et al. (2006)	**Yes**	**Yes**	No	**Yes**	No	**Yes**	No
Genter et al. (2013)	**Yes**	**Yes**	No	**Yes**	No	**Yes**	No
Genter and Stone (2014)	**Yes**	**Yes**	No	**Yes**	No	**Yes**	No
Liemhetcharat and Veloso (2014)	No	**Yes**	**Yes**	**Yes**	**Yes**	No	No
This thesis	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Learning Agents Research Group
The University of Texas at Austin

Summary of Ad Hoc Teamwork Research

- Theoretical analysis of bounds of cooperation
  - [Stone and Kraus 2010] [Agmon et al. 2014] [Chakraborty and Stone 2013]
- Selecting between hand-coded models
  - [Albrecht and Ramamoorthy 2013] [Albrecht and Ramamoorthy 2014]
- Controlling flock on agents
  - [Han et al. 2006] [Genter and Stone 2014]
- Selecting agents to form an ad hoc team
  - [Liemhetcharat and Veloso 2014]

Research Question

Research Question:

How can an agent cooperate with teammates of uncertain types on a variety of tasks?

Learning Agents Research Group
The University of Texas at Austin

Research Question

Research Question:

How can an agent cooperate with **teammates of uncertain types** on a variety of tasks?

Desiderata:

- Robustness to teammate variety

Learning Agents Research Group
The University of Texas at Austin

Research Question

Research Question:

How can an agent cooperate with teammates of uncertain types on a **variety of tasks**?

Desiderata:

- Robustness to teammate variety
- Robustness to diverse tasks

LARG
Learning Agents Research Group
The University of Texas at Austin

Research Question

Research Question:

How can an agent **cooperate** with teammates of uncertain types on a variety of tasks?

Desiderata:

- Robustness to teammate variety
- Robustness to diverse tasks
- Fast adaptation

Solution Overview

- Learn about previous teammates
- Reuse this knowledge with new teammates
- Determine which previous teammates are most similar to the new ones

Solution Overview

- Learn about previous teammates
- Reuse this knowledge with new teammates
- Determine which previous teammates are most similar to the new ones
- Planning and Learning to Adapt Swiftly to Teammates to Improve Cooperation

Solution Overview

- Learn about previous teammates
- Reuse this knowledge with new teammates
- Determine which previous teammates are most similar to the new ones
- Planning and Learning to Adapt Swiftly to Teammates to Improve Cooperation(PLASTIC)

Contributions from Proposal

- ✓ Cooperate with known teammates on a known task
- ✓ Cooperate with teammates drawn from a known set
- ✓ Cooperate with teammates **not** drawn from a known set
- ✓ Teach novice agents
- ⇒ Learn about explicit signals of teammates' intents
- ⇒ Scale to complex domains

Contributions from Proposal

- ✓ Cooperate with known teammates on a known task
- ✓ Cooperate with teammates drawn from a known set
- ✓ Cooperate with teammates **not** drawn from a known set
- ● Teach novice agents
- ✓ Learn about explicit signals of teammates' intents
- ✓ Scale to complex domains

Contributions

- PLASTIC
- Theoretical analysis
- Reasoning about communication
- TwoStageTransfer
- Empirical evaluation
- Taxonomy of ad hoc teamwork

Contributions

- ✓ PLASTIC
- Theoretical analysis
- ✓ Reasoning about communication
- ✓ TwoStageTransfer
- ✓ Empirical evaluation
- Taxonomy of ad hoc teamwork

Publications

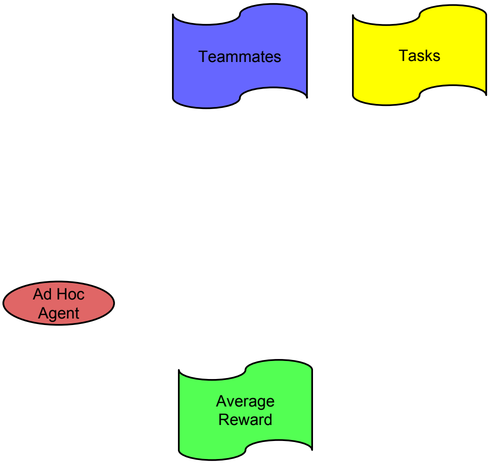
- Samuel Barrett, Peter Stone, and Sarit Kraus. Empirical evaluation of ad hoc teamwork in the pursuit domain. In *Proceedings of the Tenth International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, May 2011
- Samuel Barrett and Peter Stone. An analysis framework for ad hoc teamwork tasks. In *Proceedings of the Eleventh International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, June 2012
- Samuel Barrett, Peter Stone, Sarit Kraus, and Avi Rosenfeld. Teamwork with limited knowledge of teammates. In *Proceedings of the Twenty-Seventh Conference on Artificial Intelligence (AAAI)*, July 2013
- Samuel Barrett, Noa Agmon, Noam Hazon, Sarit Kraus, and Peter Stone. Communicating with unknown teammates. In *Proceedings of the Twenty-First European Conference on Artificial Intelligence*, August 2014
- Samuel Barrett and Peter Stone. Cooperating with unknown teammates in robot soccer. In *AAAI Workshop on Multiagent Interaction without Prior Coordination (MIPC 2014)*, July 2014

Ad Hoc Agent Evaluation

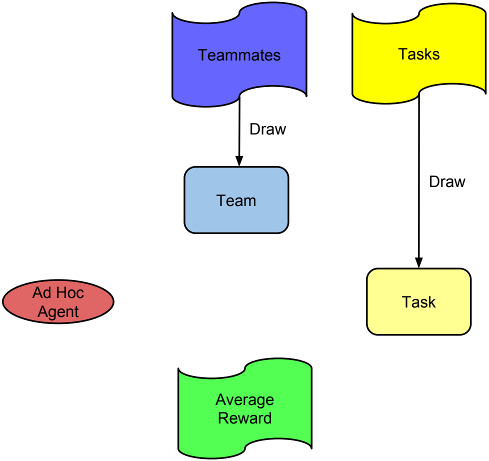
- Not whether they win, but how well they cooperate
- Depends on possible tasks
- Depends on possible teammates

[Stone et al. 2010]

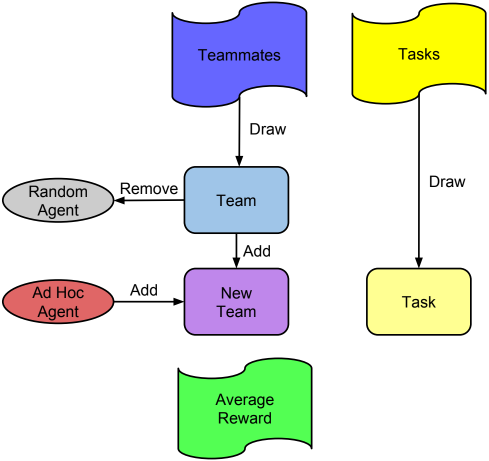
Ad Hoc Agent Evaluation



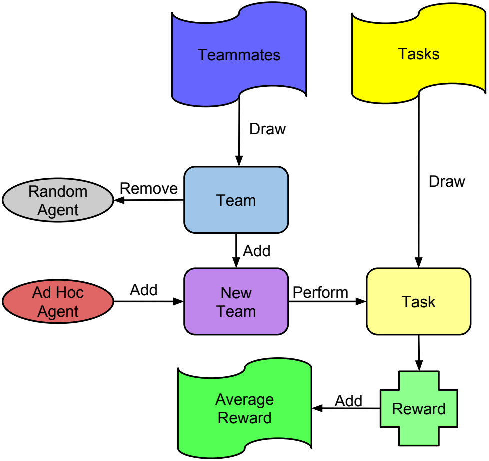
Ad Hoc Agent Evaluation



Ad Hoc Agent Evaluation



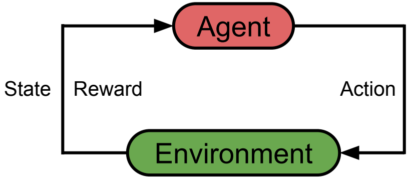
Ad Hoc Agent Evaluation



Markov Decision Process

$$\text{MDP} = \langle S, A, P, R \rangle$$

- S = State
- A = Actions
- P = transition function
- R = reward function



Methods

MDP Algorithms:

- Upper Confidence bounds for Trees (UCT)
  - Sample based planner that calculates policies for MDPs on the fly
  - [Kocsis and Szepesvari 2006]

Methods

MDP Algorithms:

- Upper Confidence bounds for Trees (UCT)
  - Sample based planner that calculates policies for MDPs on the fly
  - [Kocsis and Szepesvari 2006]
- Fitted Q iteration (FQI)
  - Sample based learning algorithm for learning a policy
  - [Ernst et al. 2005]

Methods

MDP Algorithms:

- Upper Confidence bounds for Trees (UCT)
  - Sample based planner that calculates policies for MDPs on the fly
  - [Kocsis and Szepesvari 2006]
- Fitted Q iteration (FQI)
  - Sample based learning algorithm for learning a policy
  - [Ernst et al. 2005]
- Function approximation
  - Allows generalization to nearby states
  - Tile coding: [Albus 1971]

Methods(2)

- Partially observable MDP - POMDP
  - Have to reason about which state we're in

Methods(2)

- Partially observable MDP - POMDP
  - Have to reason about which state we're in
- Partially Observable Monte Carlo Planning (POMCP)
  - Adaptation of UCT for POMDPs
  - Calculates approximate policy
  - [Silver and Veness 2010]

LARG
Learning Agents Research Group
The University of Texas at Austin

Methods(2)

- Partially observable MDP - POMDP
  - Have to reason about which state we're in
- Partially Observable Monte Carlo Planning (POMCP)
  - Adaptation of UCT for POMDPs
  - Calculates approximate policy
  - [Silver and Veness 2010]
- Decision Trees
  - Supervised learning algorithm

LARG
Learning Agents Research Group
The University of Texas at Austin

Outline

Outline

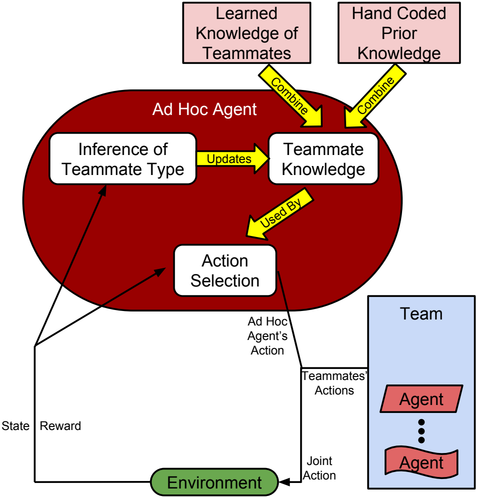
1. Introduction
2. PLASTIC
3. Results
4. Conclusion

PLASTIC Overview

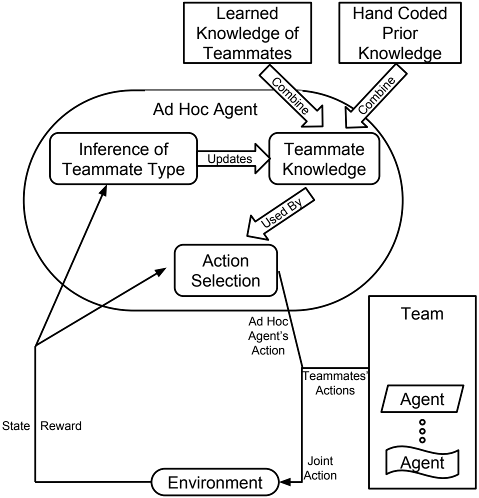
- Planning and Learning to Adapt Swiftly to Teammates to Improve Cooperation(PLASTIC)
- Learn about previous teammates
- Reuse this knowledge with new teammates
- Determine which previous teammates are most similar to the new ones

Samuel Barrett, Peter Stone, and Sarit Kraus. Empirical evaluation of ad hoc teamwork in the pursuit domain. In *Proceedings of the Tenth International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, May 2011

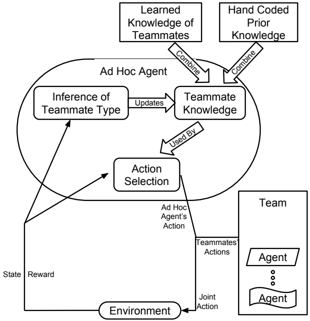
Overview of PLASTIC



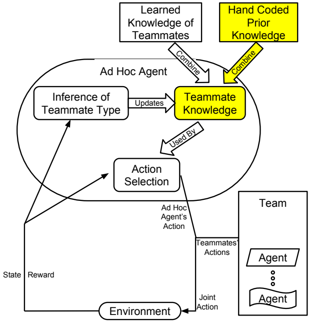
Overview of PLASTIC



Overview of PLASTIC

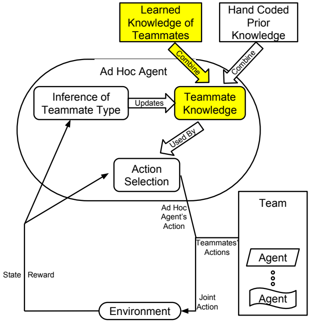


PLASTIC: Expert Knowledge



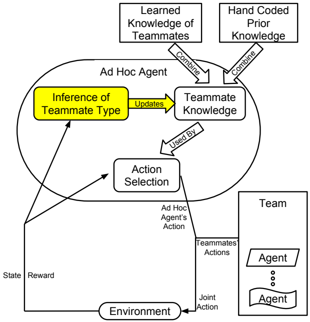
- Allow experts to provide prior knowledge
- Information about teammate behaviors or how to adapt to teammates
- Prior belief distribution over teammate behaviors

PLASTIC: Learn about Previous Teammates



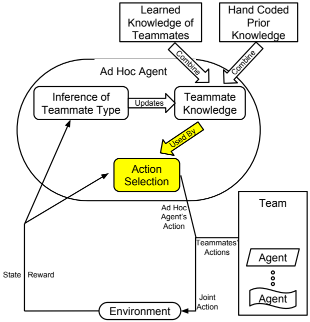
- Agent has extensive interactions with previous teammates
- Learn about previous teammates
- Use this knowledge to cooperate with new teammates

PLASTIC: Inferring Teammate Type



- Observe the actions of the teammates
- Determine the probability of a known teammate type taking the observed actions
- Update the distribution over the teammate types using a bounded loss version of Bayes' rule

PLASTIC: Action Selection



- Given the distribution over teammate types
- Given current world state
- Determine best action to take

PLASTIC–Model Motivation

- Model-based approach
- Adapts quickly to new teammates
- Reuses models of past teammates

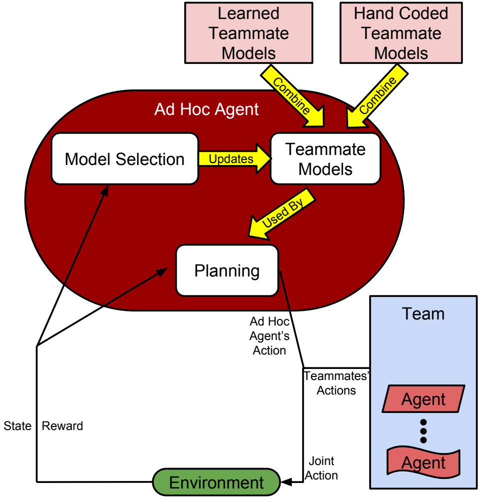
Samuel Barrett, Peter Stone, and Sarit Kraus. Empirical evaluation of ad hoc teamwork in the pursuit domain. In *Proceedings of the Tenth International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, May 2011

PLASTIC–Model Motivation

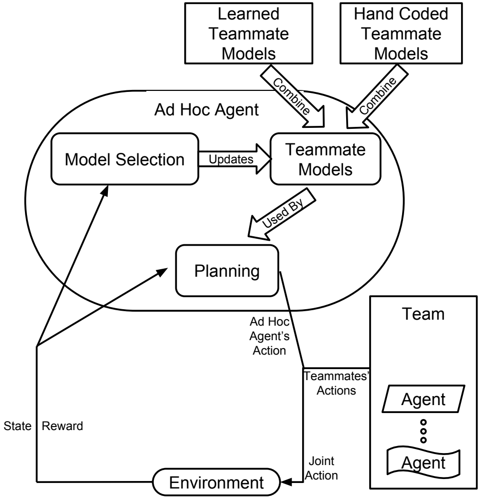
- Model-based approach
- Adapts quickly to new teammates
- Reuses models of past teammates
- Given the true model of the environment and teammates, can calculate the optimal policy
- Can select actions given a distribution over model types

Samuel Barrett, Peter Stone, and Sarit Kraus. Empirical evaluation of ad hoc teamwork in the pursuit domain. In *Proceedings of the Tenth International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, May 2011

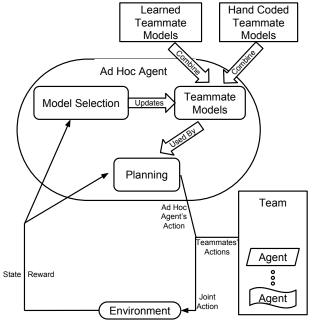
Overview of PLASTIC–Model



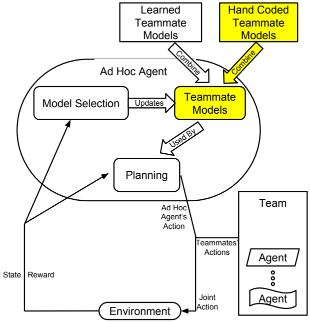
Overview of PLASTIC–Model



Overview of PLASTIC–Model

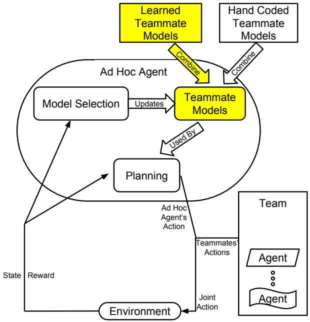


PLASTIC–Model: Expert Knowledge



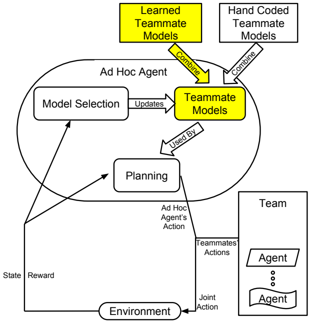
- Model-based approach
- Expert provides teammate models
- Hand-coded behaviors of potential teammates

PLASTIC–Model: Learn about Previous Teammates



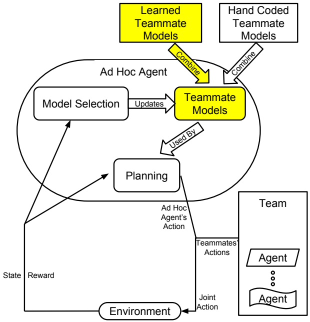
- Collect samples of past teammates
- Mapping from states to actions

PLASTIC–Model: Learn about Previous Teammates



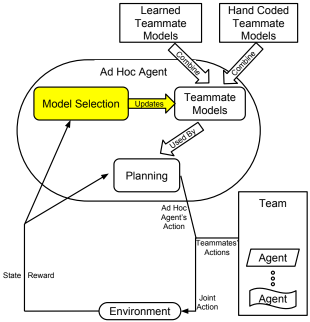
- Collect samples of past teammates
- Mapping from states to actions
- Supervised learning problem
- Use existing learning algorithms, such as decision trees

PLASTIC–Model: Learn about Previous Teammates



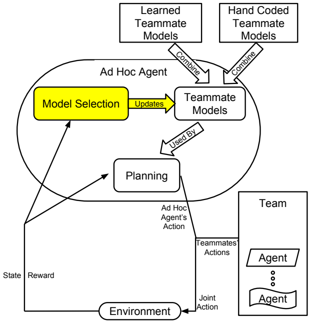
- Collect samples of past teammates
- Mapping from states to actions
- Supervised learning problem
- Use existing learning algorithms, such as decision trees
- Can use transfer learning, such as TwoStageTransfer

PLASTIC–Model: Inferring Teammate Type



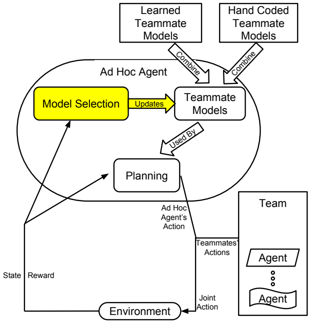
- Update models using observed actions
- Use Bayes' rule

PLASTIC–Model: Inferring Teammate Type



- Update models using observed actions
- Use Bayes' rule
- But may have some bad predictions
- Use bounded loss

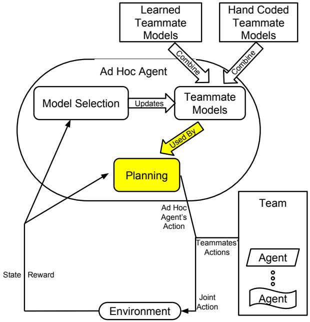
PLASTIC–Model: Inferring Teammate Type



- Update models using observed actions
- Use Bayes' rule
- But may have some bad predictions
- Use bounded loss

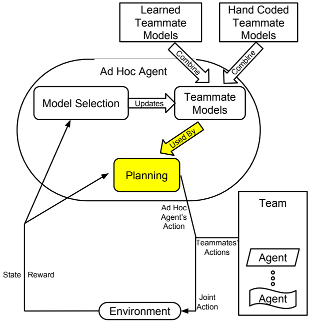
$$\text{loss} = 1 - P(\text{actions}|\text{model})$$
$$P(\text{model}|\text{actions}) \propto (1 - \eta * \text{loss}) * P(\text{model})$$

PLASTIC–Model: Action Selection



- Select best action
- Know model
- Know distribution over teammates

PLASTIC–Model: Action Selection



- Select best action
- Know model
- Know distribution over teammates
- Solve using MDP planners, such as UCT

PLASTIC–Policy Motivation

- Policy-based approach
- Reuses policies for cooperating with past teammates
- Adapts quickly to new teammates

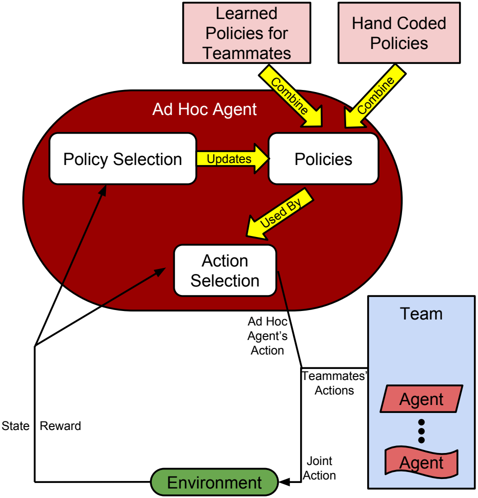
Samuel Barrett and Peter Stone. Cooperating with unknown teammates in robot soccer. In *AAAI Workshop on Multiagent Interaction without Prior Coordination (MIPC 2014)*, July 2014

PLASTIC–Policy Motivation

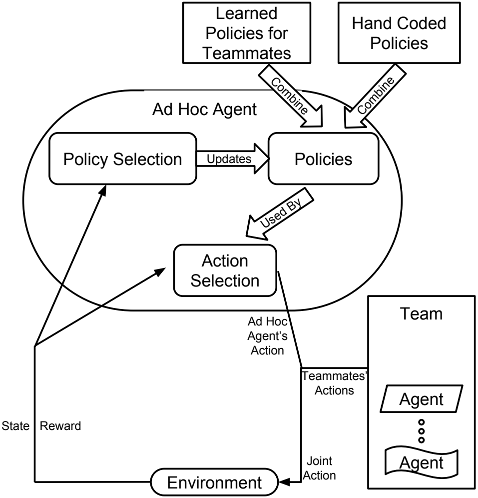
- Policy-based approach
- Reuses policies for cooperating with past teammates
- Adapts quickly to new teammates
- Policy-based methods better handle complex, noisy domains
  - Many robotic tasks have been better solved using policy-based approaches
- Fast online computation

Samuel Barrett and Peter Stone. Cooperating with unknown teammates in robot soccer. In *AAAI Workshop on Multiagent Interaction without Prior Coordination (MIPC 2014)*, July 2014

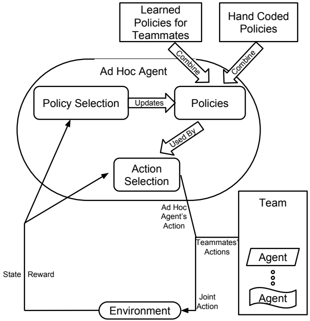
Overview of PLASTIC–Policy



Overview of PLASTIC–Policy

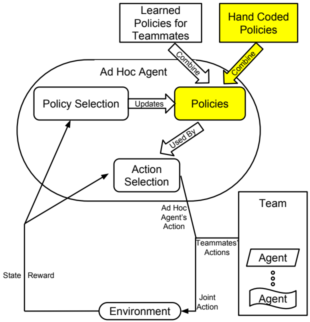


Overview of PLASTIC–Policy



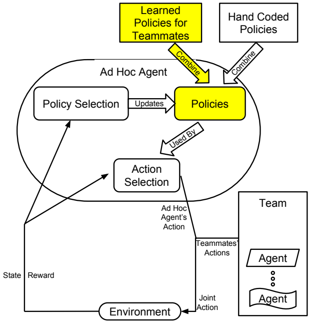
LARG Learning Agents Research Group
The University of Texas at Austin

PLASTIC–Policy: Expert Knowledge



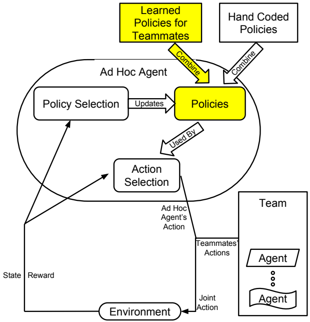
- Policy-based approach
- Expert provides policies for cooperating with teammates
- Hand-coded policies for behaving intelligently

PLASTIC–Policy: Learn about Previous Teammates



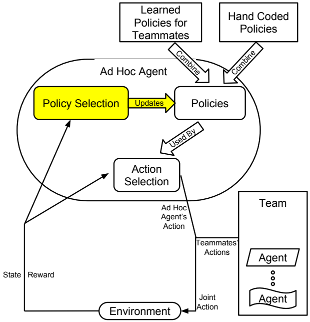
- Collect samples of past teammates
- $\langle s, a, r, s' \rangle$

PLASTIC–Policy: Learn about Previous Teammates



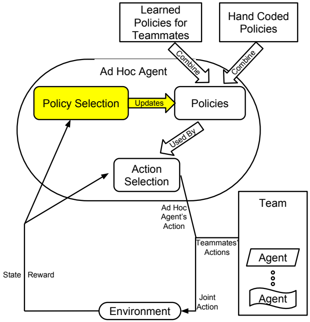
- Collect samples of past teammates
- $\langle s, a, r, s' \rangle$
- Use existing policy learning algorithms, such as fitted Q iteration

PLASTIC–Policy: Inferring Teammate Type



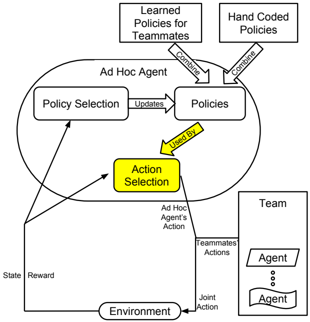
- As in PLASTIC–Model
- Update models using observed actions
- Use Bayes' rule with bounded loss
- But do not have full model

PLASTIC–Policy: Inferring Teammate Type



- As in PLASTIC–Model
- Update models using observed actions
- Use Bayes' rule with bounded loss
- But do not have full model
- Estimate using a nearest neighbors transition function

PLASTIC–Policy: Action Selection



- Straight-forward
- Use policy with highest probability
- Select best action for policy

Outline

Dimensions

Team Knowledge: Does the ad hoc agent know what its teammates' actions will be for a given state, before interacting with them?

Samuel Barrett and Peter Stone. An analysis framework for ad hoc teamwork tasks. In *Proceedings of the Eleventh International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, June 2012

Dimensions

Team Knowledge: Does the ad hoc agent know what its teammates' actions will be for a given state, before interacting with them?

Environment Knowledge: Does the ad hoc agent know the transition and reward distribution given a joint action and state before interacting with the environment?

Samuel Barrett and Peter Stone. An analysis framework for ad hoc teamwork tasks. In *Proceedings of the Eleventh International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, June 2012

Dimensions

Team Knowledge: Does the ad hoc agent know what its teammates' actions will be for a given state, before interacting with them?

Environment Knowledge: Does the ad hoc agent know the transition and reward distribution given a joint action and state before interacting with the environment?

Reactivity of teammates: How much does the ad hoc agent's actions affect those of its teammates?

Samuel Barrett and Peter Stone. An analysis framework for ad hoc teamwork tasks. In *Proceedings of the Eleventh International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, June 2012

LARG
Learning Agents Research Group
The University of Texas at Austin

Overview of Empirical Results

- Test the hypothesis that PLASTIC enables agents to quickly adapt to new teammates in a variety of possible ad hoc teamwork scenarios

LARG
Learning Agents Research Group
The University of Texas at Austin

Overview of Empirical Results

- Test the hypothesis that PLASTIC enables agents to quickly adapt to new teammates in a variety of possible ad hoc teamwork scenarios
- Test in 3 domains: Bandit, Pursuit, and HFO (simulated soccer)

Overview of Empirical Results

- Test the hypothesis that PLASTIC enables agents to quickly adapt to new teammates in a variety of possible ad hoc teamwork scenarios
- Test in 3 domains: Bandit, Pursuit, and HFO (simulated soccer)
- Can PLASTIC help when there is limited communication?
- Can PLASTIC learn models of teammates?
- Is TwoStageTransfer effective for transferring knowledge of past teammates?
- Can PLASTIC scale to complex domains?

Overview of Experiments

Domain	Teammate Type	Teammate Knowledge	Teammates Previously Seen	Environment Known	Number of Teammates	Uses Comm.	Continuous State/Actions	PLASTIC–Model or PLASTIC–Policy
Bandit	HC	**Param.** HC Set	Yes	Yes	7	**Yes**	No	Model
Bandit	**Ext.**	**Param.** HC Set	**No**	Yes	1–9	**Yes**	No	Model
Bandit	HC and **Ext.**	**Param.** HC Set	Yes and **No**	**No**	1–9	**Yes**	No	Model
Pursuit	HC	Known	Yes	Yes	3	No	No	Model
Pursuit	HC	HC Set	Yes	Yes	3	No	No	Model
Pursuit	**Ext.**	HC Set	**No**	Yes	3	No	No	Model
Pursuit	**Ext.**	Learned Set	Yes and **No**	Yes	3	No	No	Model
Pursuit	**Ext.**	Learned Set + **TwoStageTransfer**	**Briefly**	Yes	3	No	No	Model
Limited HFO	**Ext.**	Learned Set	Yes	Yes	1	No	**Yes**	Policy
Full HFO	**Ext.**	Learned Set	Yes	Yes	3	No	**Yes**	Policy

Multi-armed bandit

- Bernoulli arms
- Multiagent: each agent pulls an arm
  - Ad hoc agent observes all payoffs
  - Other agents observe their own

Multi-armed bandit

- Bernoulli arms
- Multiagent: each agent pulls an arm
  - Ad hoc agent observes all payoffs
  - Other agents observe their own
- Limited communication
  - Fixed set of messages
  - Has explicit cost

Multi-armed bandit

- Bernoulli arms
- Multiagent: each agent pulls an arm
  - Ad hoc agent observes all payoffs
  - Other agents observe their own
- Limited communication
  - Fixed set of messages
  - Has explicit cost
- Goal: Maximize payoffs and minimize communication costs

Learning Agents Research Group
The University of Texas at Austin

Bandit Domain: Communication

- **Last observation** - last arm chosen and resulting payoff
- **Arm mean** - mean and number of pulls of one arm
- **Suggestion** - suggest that teammates should pull an arm

Bandit Domain: Teammates

- Tightly coordinated

Bandit Domain: Teammates

- Tightly coordinated
  - Team shares knowledge through communication
  - Do **not** need to track each agent's pulls

Bandit Domain: Teammates

- Hand-Coded
  - ε-Greedy – mostly greedy with chance of exploration
  - UCB(c) – selects greedily with respect to upper confidence bounds
  - Have probability of following suggestion sent by the ad hoc agent

Bandit Domain: Teammates

- Hand-Coded
  - ε-Greedy – mostly greedy with chance of exploration
  - UCB(c) – selects greedily with respect to upper confidence bounds
  - Have probability of following suggestion sent by the ad hoc agent
- Externally-created
  - Created by students for project
  - Not tightly coordinated
  - Not considering ad hoc teamwork

Theoretical Analysis: Setup

- 2 arms
- Cooperating with hand-coded teammates
- PLASTIC–Model given set of hand-coded models
  - Parameterized set
- Unknown arm payoffs

Samuel Barrett, Noa Agmon, Noam Hazon, Sarit Kraus, and Peter Stone. Communicating with unknown teammates. In *Proceedings of the Twenty-First European Conference on Artificial Intelligence*, August 2014

LARG Learning Agents Research Group The University of Texas at Austin

Theoretical Analysis: Setup

- Model problem as POMDP
  - Teammate behavior type and arm distributions is partially observable
  - Team is tightly coordinated
  - States are the team's pulls and successes

Theoretical Analysis: Setup

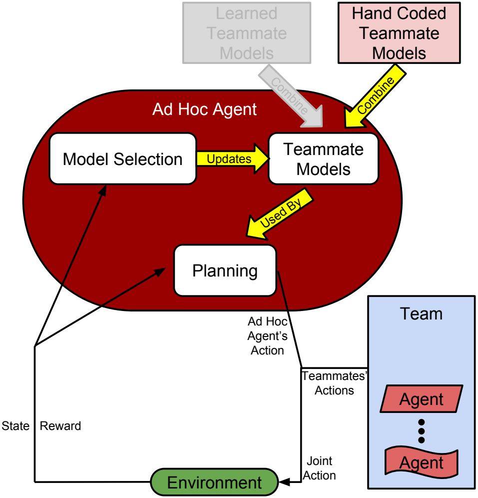
- Model problem as POMDP
  - Teammate behavior type and arm distributions is partially observable
  - Team is tightly coordinated
  - States are the team's pulls and successes
- Bound the size of the belief space about teammates' behaviors and the arms' distributions
- Proves that the POMDP can be approximately solved in polynomial time

Bandit Methods

- Model as a POMDP
- Apply PLASTIC–Model
- Approximate the planning and belief updates using Partially Observable Monte Carlo Planning (POMCP)

Samuel Barrett, Noa Agmon, Noam Hazon, Sarit Kraus, and Peter Stone. Communicating with unknown teammates. In *Proceedings of the Twenty-First European Conference on Artificial Intelligence*, August 2014

Overview of PLASTIC–Model



Potential Behaviors

- **Match** – Plays as if it were another agent of the team's type, but can observe all agents' results
- **NoComm** – Pulls the best arm and does not communicate
- **Obs** – Pulls the best arm and sends its last observation
- **PLASTIC–Model** – Selects arms and messages using PLASTIC–Model

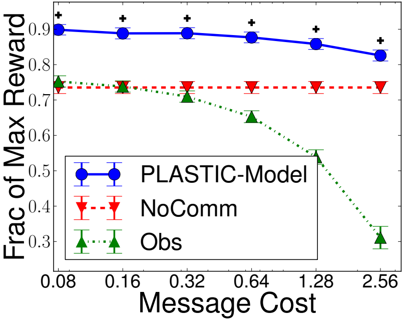
LARG
Learning Agents Research Group
The University of Texas at Austin

Experimental Setup

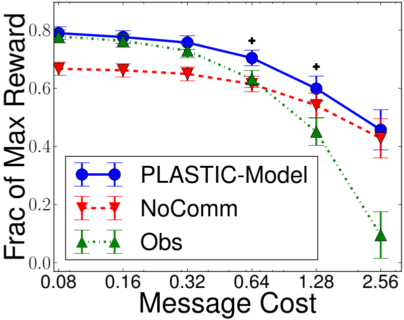
- Ad hoc agent expects ε-greedy or UCB(c) teammates
- 100 trials, 10 rounds, 7 teammates, 3 arms
- Message costs randomly chosen and depends on the amount of information
  - Mean is highest, obs is middle, and sugg is lowest
- Statistical tests via Wilcoxon signed-rank test with $p < 0.05$

Known Arms

Known Arms

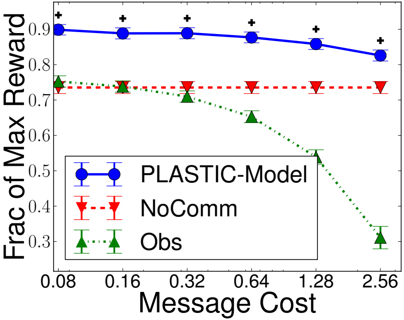


ε-greedy teammates

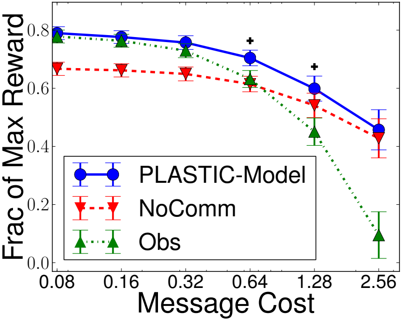


Externally-created teammates

Known Arms



ε-greedy teammates



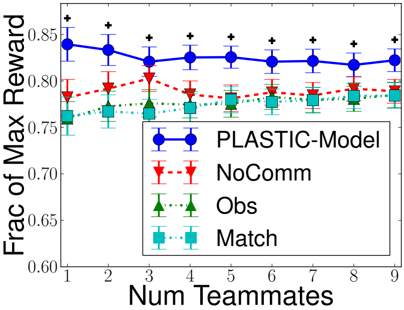
Externally-created teammates

- PLASTIC–Model outperforms other approaches
- PLASTIC–Model scales as the message costs rise

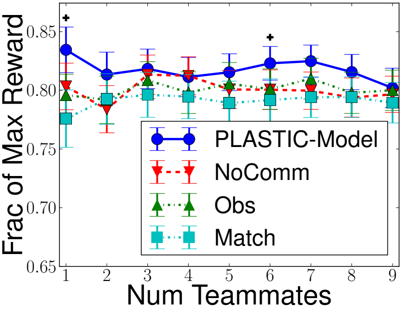
Unknown Arms

- Ad hoc agent does not know the true arm payoffs
- Ad hoc agent must balance learning about the environment, learning about its teammates, and exploiting its current knowledge

Unknown Arms

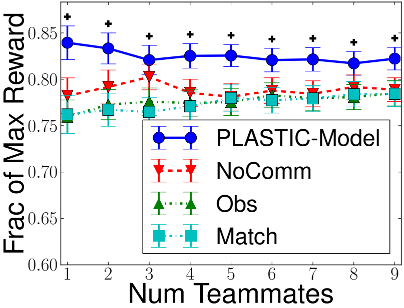


Hand-coded teammates

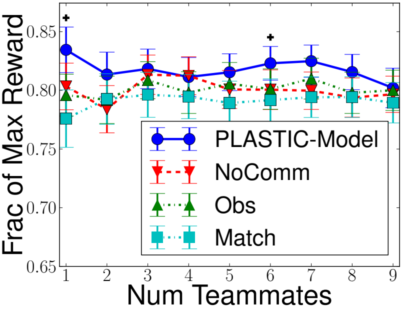


Externally-created teammates

Unknown Arms



Hand-coded teammates



Externally-created teammates

- PLASTIC–Model outperforms other approaches

Summary of Bandit Experiments

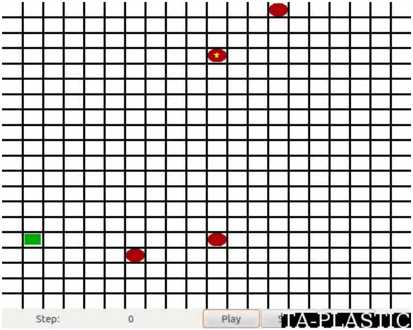
- Evaluated PLASTIC–Model in the bandit domain
- PLASTIC–Model can effectively reason about limited communication
- Cooperated successfully with hand-coded teammates
- Cooperated successfully with externally-created teammates even though its prior knowledge was incorrect
- Performed well when environment was unknown

Pursuit Domain

- Grid world - Torus
- 4 predators try to surround prey
- Act simultaneously
- Collisions randomly decided - loser stays still

[Barrett et al. 2011]

[Barrett et al. 2013]



Pursuit Domain: Agent Control

- Observe positions of all agents
- Cannot explicitly communicate
- 5 actions: Stay still, up, down, left, and right
- Prey acts randomly

Pursuit Domain: Teammates

- Hand-coded
  - 4 types created by me

Pursuit Domain: Teammates

- Hand-coded
  - 4 types created by me
- Externally-created
  - Created by students
  - Designed to cooperate with agents from same student
  - Student - 29 from students

Pursuit Methods

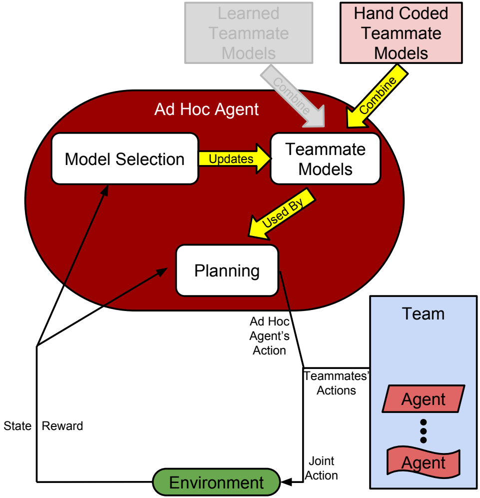
- Known environment
- Apply PLASTIC–Model
- Plan using Upper Confidence bounds for Trees (UCT)

Experimental Setup

- Number of captures in 500 steps
  - After a capture, the prey is randomly reset
- 1,000 trials

- Statistical tests via Wilcoxon signed-rank test with $p < 0.01$

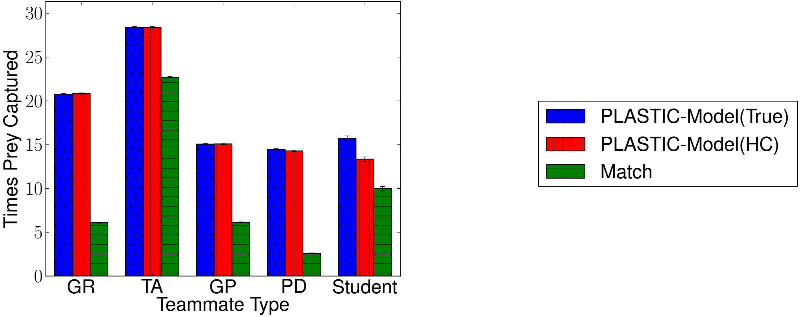
Overview of PLASTIC–Model



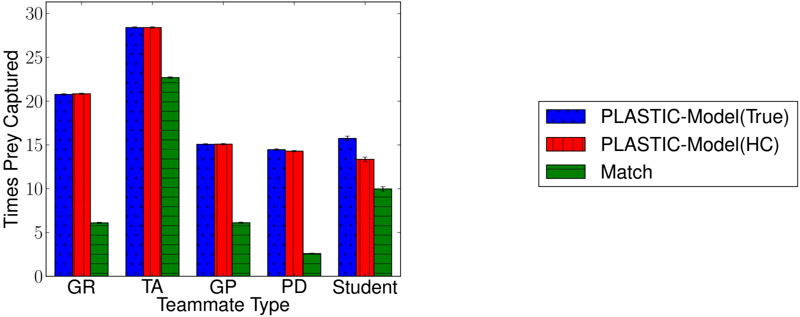
Potential Behaviors

- **Match** – Plays as if it were another agent of the team's type
- **PLASTIC–Model(True)** – Given the current teammates' true behavior
- **PLASTIC–Model(HC)** – Given the 4 hand-coded models

Hand-coded Knowledge



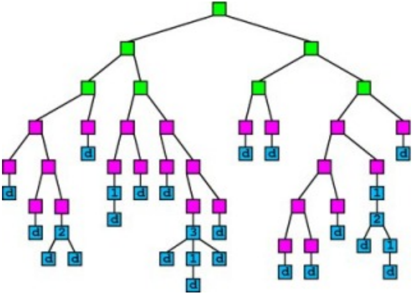
Hand-coded Knowledge



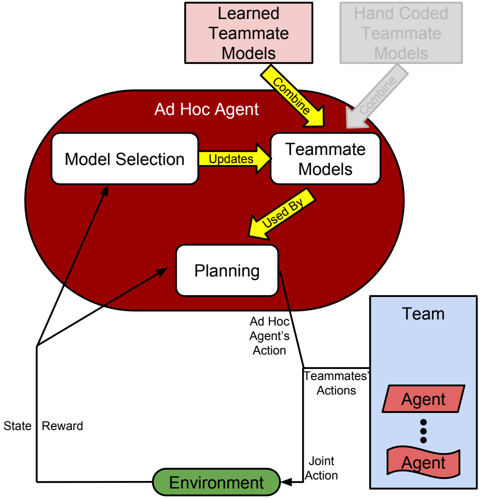
- PLASTIC–Model outperforms matching
- Selecting from a set of models performs well, even when models are incorrect

Learning about Teammates

- Learn mapping from world state to teammates' actions
- Learn one model per team
- Decision tree



Overview of PLASTIC–Model



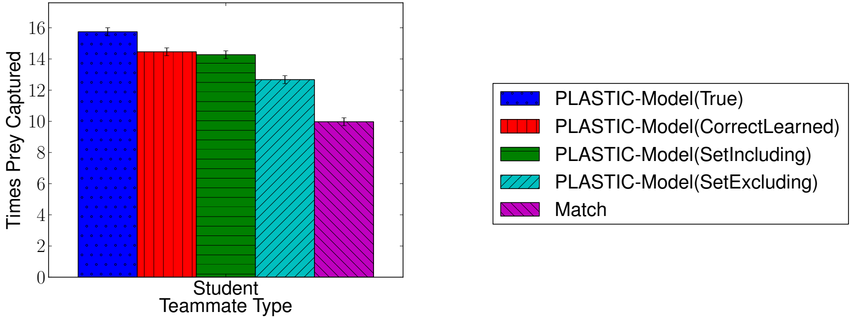
Learning about Teammates

- **Match** – Plays as if it were another agent of the team's type
- **PLASTIC–Model(True)** – Given the current teammates' true behavior

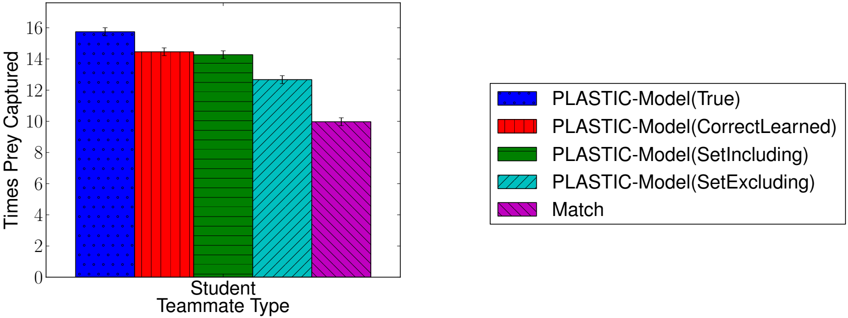
Learning about Teammates

- **Match** – Plays as if it were another agent of the team's type
- **PLASTIC–Model(True)** – Given the current teammates' true behavior
- **PLASTIC–Model(CorrectLearned)** – Given the decision tree learned from the current teammates
- **PLASTIC–Model(SetIncluding)** – Given decision trees for **all** 29 teammates
- **PLASTIC–Model(SetExcluding)** – Given decision trees for the **other** 28 teammates

Learning About Teammates



Learning About Teammates



- PLASTIC–Model outperforms matching the teammates' behavior
- Learned models generalize to previously unseen teammates

Teammates with limited observations

- Few observations of current teammate type
- Many observations of other teammate types

Teammates with limited observations

- Few observations of current teammate type
- Many observations of other teammate types
- Transfer learning problem

Teammates with limited observations

- Few observations of current teammate type
- Many observations of other teammate types
- Transfer learning problem
- Use the information that the prior experiences come from different sources
- Consider past teammates' similarities separately

TwoStageTransfer Overview

- Transfer data, not models
- Combine all data, but weight data coming from different teammates differently
- Find best weighting of data from prior teammates
- Test weightings with cross validation

Samuel Barrett, Peter Stone, Sarit Kraus, and Avi Rosenfeld. Teamwork with limited knowledge of teammates. In *Proceedings of the Twenty-Seventh Conference on Artificial Intelligence (AAAI)*, July 2013

TwoStageTransfer Description

- Find best weighting of data from each past teammate

Samuel Barrett, Peter Stone, Sarit Kraus, and Avi Rosenfeld. Teamwork with limited knowledge of teammates. In *Proceedings of the Twenty-Seventh Conference on Artificial Intelligence (AAAI)*, July 2013

TwoStageTransfer Description

- Find best weighting of data from each past teammate
  - For n past teammates and m weightings
  - Checking all possible weightings is m^n
  - TwoStageTransfer checks $nm + nm = 2nm$ weightings

Samuel Barrett, Peter Stone, Sarit Kraus, and Avi Rosenfeld. Teamwork with limited knowledge of teammates. In *Proceedings of the Twenty-Seventh Conference on Artificial Intelligence (AAAI)*, July 2013

TwoStageTransfer Description

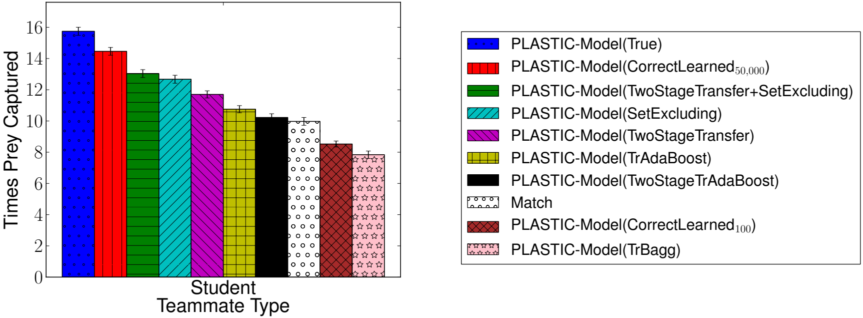
- Find best weighting of data from each past teammate
  - For n past teammates and m weightings
  - Checking all possible weightings is m^n
  - TwoStageTransfer checks $nm + nm = 2nm$ weightings
- Greedily choose past teammates ordered by improvement with current teammate
- Search over weighting of past teammate's data

Samuel Barrett, Peter Stone, Sarit Kraus, and Avi Rosenfeld. Teamwork with limited knowledge of teammates. In *Proceedings of the Twenty-Seventh Conference on Artificial Intelligence (AAAI)*, July 2013

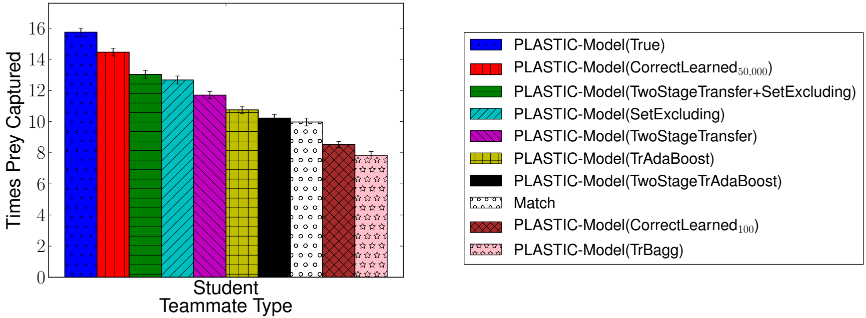
TwoStageTransfer Advantages

- Uses the information that the prior experiences come from different sources
- Can use all prior experiences
- Considers past teammate weights separately
- Efficient

Teammates with Limited Observations



Teammates with Limited Observations



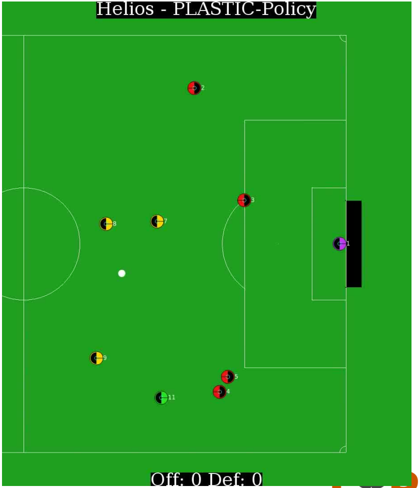
- TwoStageTransfer effectively transfers knowledge from past teammates
- Combining the models from past teammates and the model built using TwoStageTransfer performs best

Summary of Pursuit Experiments

- Evaluated PLASTIC–Model in the pursuit domain
- PLASTIC–Model can handle multiagent coordination
- Cooperated successfully with hand-coded teammates
- Cooperated successfully with externally-created teammates
- Can cooperate with previously unseen teammates
- Can learn models of teammates
- TwoStageTransfer performs well for learning models of new teammates with few observations

Half Field Offense

- Complex observations and actuators
- Offense tries to score
- Episode ends when:
  - Score
  - Ball leaves half field
  - Ball captured by defense

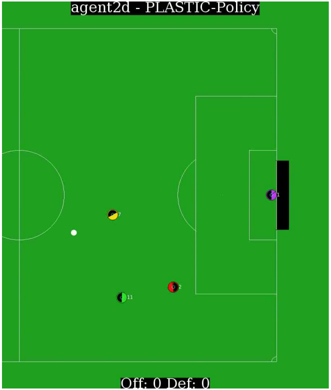


[Kalyanakrishnan et al. 2007]

HFO: Versions

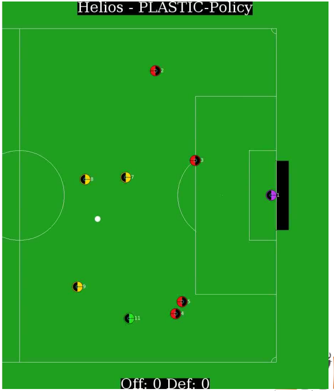
Limited

- 2 offensive players
- 2 defensive players



Full

- 4 offensive players
- 5 defensive players



HFO: Teammates

- Externally-created teammates
- Total of 7 teammate types

HFO: Teammates

- Externally-created teammates
- Total of 7 teammate types
- 6 of top 8 teams from the 2013 competition
  - aut
  - axiom
  - cyrus
  - gliders
  - helios
  - yushan
- Plus the agent2d code release

Learning Agents Research Group
The University of Texas at Austin

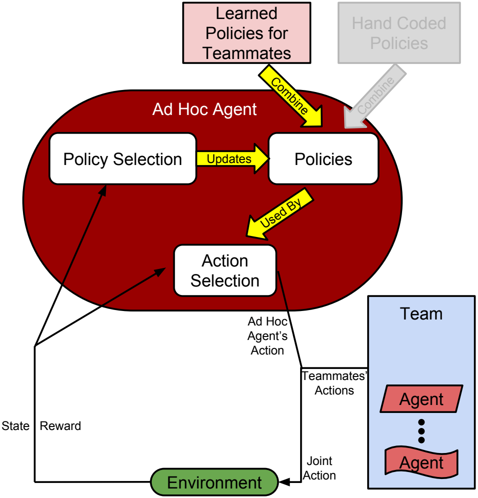
HFO Methods

- Complex, noisy domain
- Learning a model is difficult

HFO Methods

- Complex, noisy domain
- Learning a model is difficult
- Apply PLASTIC–Policy
- Learn policies using Fitted Q Iteration (FQI) with CMAC tile coding
- Select between policies using bounded loss version of Bayes' rule

Overview of PLASTIC–Policy

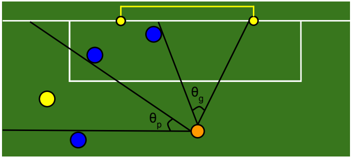


Applying PLASTIC–Policy

- Continuous state space
- Continuous actions

Continuous State Space

- Agent's x,y position and orientation
- Agent's goal opening angle
- Teammate's goal opening angle
- Distance to opponent
- Distance from teammate to opponent
- Pass opening angle
- Distance to teammate



Continuous Actions

- Use high level actions

- 6 with ball:
  - Shoot
  - Short dribble
  - Long dribble
  - Pass_0
  - Pass_1
  - Pass_2
- 7 with**out** ball:
  - Stay still
  - Towards the ball
  - Towards the opposing goal
  - Towards the nearest teammate
  - Away from the nearest teammate
  - Towards the nearest opponent
  - Away from the nearest opponent

Approaches

- Given observations of past teammates
- **Combined Policy** – combine observations from all teams to learn one policy

Approaches

- Given observations of past teammates
- **Combined Policy** – combine observations from all teams to learn one policy
- Learn 1 policy for each past team

Approaches

- Given observations of past teammates
- **Combined Policy** – combine observations from all teams to learn one policy
- Learn 1 policy for each past team
  - **Bandit** – selects policies using a bandit-based approach
    - Pulling an arm = 1 game of HFO

Approaches

- Given observations of past teammates
- **Combined Policy** – combine observations from all teams to learn one policy
- Learn 1 policy for each past team
  - **Bandit** – selects policies using a bandit-based approach
    - Pulling an arm = 1 game of HFO
  - **PLASTIC–Policy** – selects policies using bounded loss version of Bayesian update
    - Update probabilities of policies after each action

Learning Agents Research Group
The University of Texas at Austin

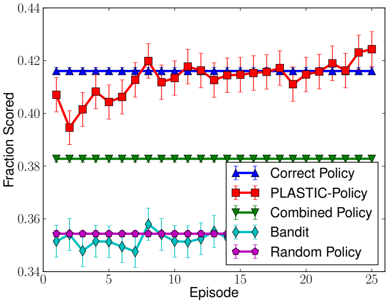
Approaches

- Given observations of past teammates
- **Combined Policy** – combine observations from all teams to learn one policy
- Learn 1 policy for each past team
  - **Bandit** – selects policies using a bandit-based approach
    - Pulling an arm = 1 game of HFO
  - **PLASTIC–Policy** – selects policies using bounded loss version of Bayesian update
    - Update probabilities of policies after each action
  - **Correct Policy** – uses the policy learned for the current teammates
  - **Random Policy** – selects a random policy

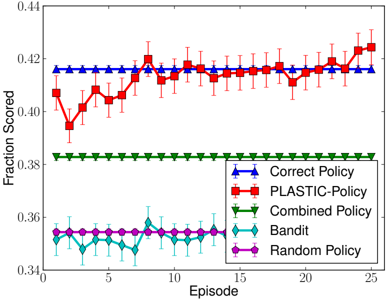
Experimental Setup

- 1,000 trials
- Fraction of trials that offense scores

Limited HFO

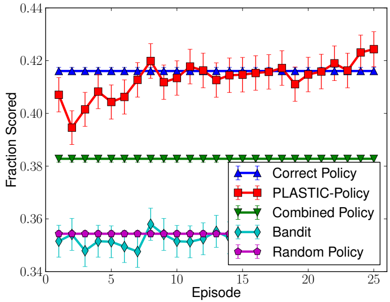

Fraction Scored
Episode
Correct Policy
PLASTIC-Policy
Combined Policy
Bandit
Random Policy

Limited HFO



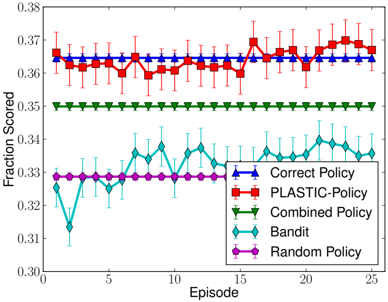
- Bandit reaches 0.382 after 1,750 episodes and 0.418 after 10,000 episodes

Limited HFO

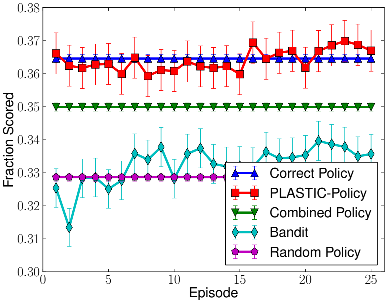


- Bandit reaches 0.382 after 1,750 episodes and 0.418 after 10,000 episodes
- PLASTIC–Policy outperforms Combined Policy – naively ignoring the teammate types
- PLASTIC–Policy outperforms the Bandit approach – Bayesian updates converge much faster

Full HFO

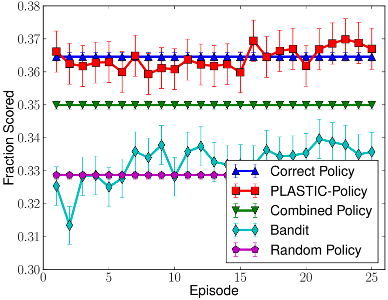

Fraction Scored
0.38
0.37
0.36
0.35
0.34
0.33
0.32
0.31
0.30
0
5
10
15
20
25
Episode
Correct Policy
PLASTIC-Policy
Combined Policy
Bandit
Random Policy

Full HFO



- Bandit reaches 0.350 after 12,000 episodes and 0.357 after 20,000 episodes

Full HFO



- Bandit reaches 0.350 after 12,000 episodes and 0.357 after 20,000 episodes
- PLASTIC–Policy outperforms Combined Policy – naively ignoring the teammate types
- PLASTIC–Policy outperforms the Bandit approach – Bayesian updates converge much faster

Summary of HFO Experiments

- PLASTIC–Policy is effective in a complex domain with continuous states and actions
- PLASTIC–Policy can cooperate with externally-created teammates from the 2013 RoboCup competition
  - Previous tests used agents created by students
  - These agents were created over years of effort
- Using the bounded loss Bayesian updates outperforms bandit approaches
- Learning specialized policies for each teammate outperforms using a single policy for all agents

Outline

Future Work Overview

- Robotic tasks
- Learning about the environment
- Human interactions
- Improvements in transfer learning

Robotic Tasks

- Scaling to larger, noisier domains
- Policy search is more promising
- Determining teammate types may be more difficult due to observation noise
- Possibly address by reasoning about value of information

Learning about the Environment

- Balance exploring the teammates, exploring the environment, and exploiting current knowledge
- Increases complexity to POMDP
- Handled partially in the bandit setting
- Initially, consider domains drawn from a limited set
  - Can be handle by similar model selection approaches

Human Interactions

- Not inherently different
- Limited trials
- Noisier behaviors
- May need to cluster teammates' behaviors for learning

Improvements in Transfer Learning

- Consider transferring from specific parts of the state space more
  - May be handled similarly to selecting a source set in TwoStageTransfer
  - May be able to use a hierarchical Bayesian model
- Transfer learning in policy-based approaches

Contributions

- PLASTIC
- Theoretical analysis
- Reasoning about communication
- TwoStageTransfer
- Empirical evaluation
- Taxonomy of ad hoc teamwork

Contributions

- PLASTIC
  - Reuses knowledge about previous teammates
  - Determines which previous teammates best match the current teammates
- Theoretical analysis
- Reasoning about communication
- TwoStageTransfer
- Empirical evaluation
- Taxonomy of ad hoc teamwork

Contributions

- PLASTIC
- Theoretical analysis
  - Proves PLASTIC is computationally tractable in the bandit domain
- Reasoning about communication
- TwoStageTransfer
- Empirical evaluation
- Taxonomy of ad hoc teamwork

Contributions

- PLASTIC
- Theoretical analysis
- Reasoning about communication
  - PLASTIC can plan to act effectively in domains with limited communication
- TwoStageTransfer
- Empirical evaluation
- Taxonomy of ad hoc teamwork

Contributions

- PLASTIC
- Theoretical analysis
- Reasoning about communication
- TwoStageTransfer
  - Allows efficient transfer of knowledge from many past teammates
- Empirical evaluation
- Taxonomy of ad hoc teamwork

Contributions

- PLASTIC
- Theoretical analysis
- Reasoning about communication
- TwoStageTransfer
- Empirical evaluation
  - Results in bandit, pursuit, and HFO domains
  - Show that PLASTIC handles communication, coordination, and complex tasks
- Taxonomy of ad hoc teamwork

Contributions

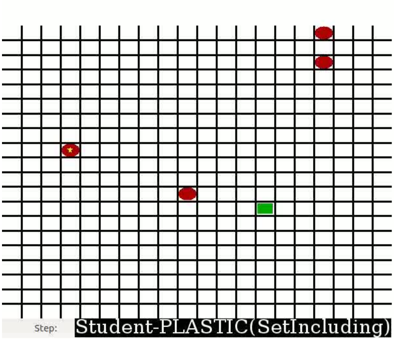
- PLASTIC
- Theoretical analysis
- Reasoning about communication
- TwoStageTransfer
- Empirical evaluation
- Taxonomy of ad hoc teamwork
  - Identifies dimensions for describing ad hoc teamwork problems

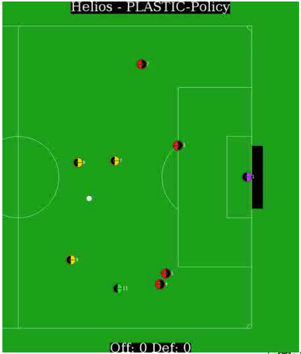
Publications

- Samuel Barrett, Peter Stone, and Sarit Kraus. Empirical evaluation of ad hoc teamwork in the pursuit domain. In *Proceedings of the Tenth International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, May 2011
- Samuel Barrett and Peter Stone. An analysis framework for ad hoc teamwork tasks. In *Proceedings of the Eleventh International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, June 2012
- Samuel Barrett, Peter Stone, Sarit Kraus, and Avi Rosenfeld. Teamwork with limited knowledge of teammates. In *Proceedings of the Twenty-Seventh Conference on Artificial Intelligence (AAAI)*, July 2013
- Samuel Barrett, Noa Agmon, Noam Hazon, Sarit Kraus, and Peter Stone. Communicating with unknown teammates. In *Proceedings of the Twenty-First European Conference on Artificial Intelligence*, August 2014
- Samuel Barrett and Peter Stone. Cooperating with unknown teammates in robot soccer. In *AAAI Workshop on Multiagent Interaction without Prior Coordination (MIPC 2014)*, July 2014

Thank You!

Agents can quickly adapt to unknown teammates by transferring knowledge learned from previous teammates.

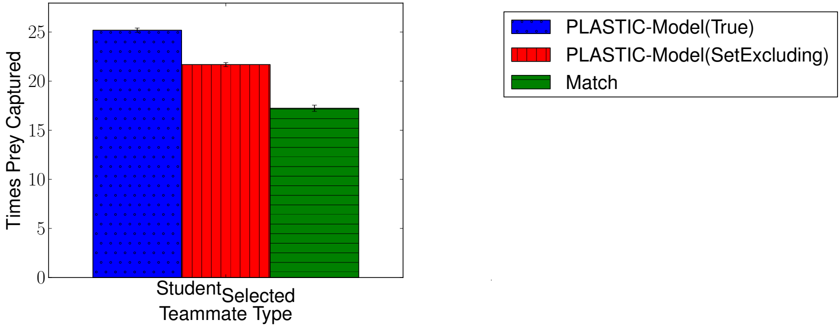




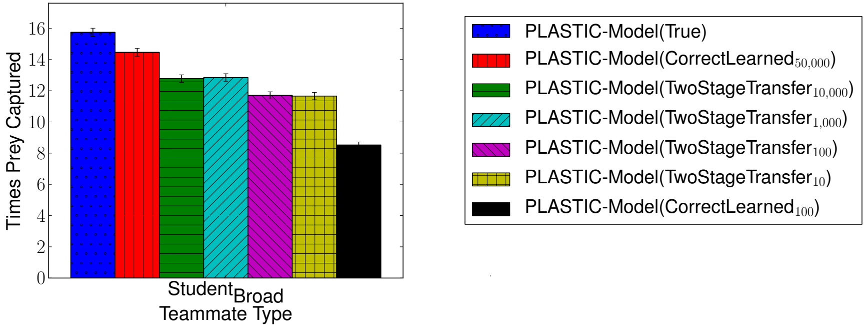
Learning Agents Research Group
The University of Texas at Austin

– This slide is intentionally left blank. –

Varying amounts of target data



Varying amounts of target data



– This slide is intentionally left blank. –