Revisiting Asimov's First Law: A Response to the Call to Arms

David V. Pynadath, Milind Tambe
USC Information Sciences Institute
{pynadath,tambe}@isi.edu

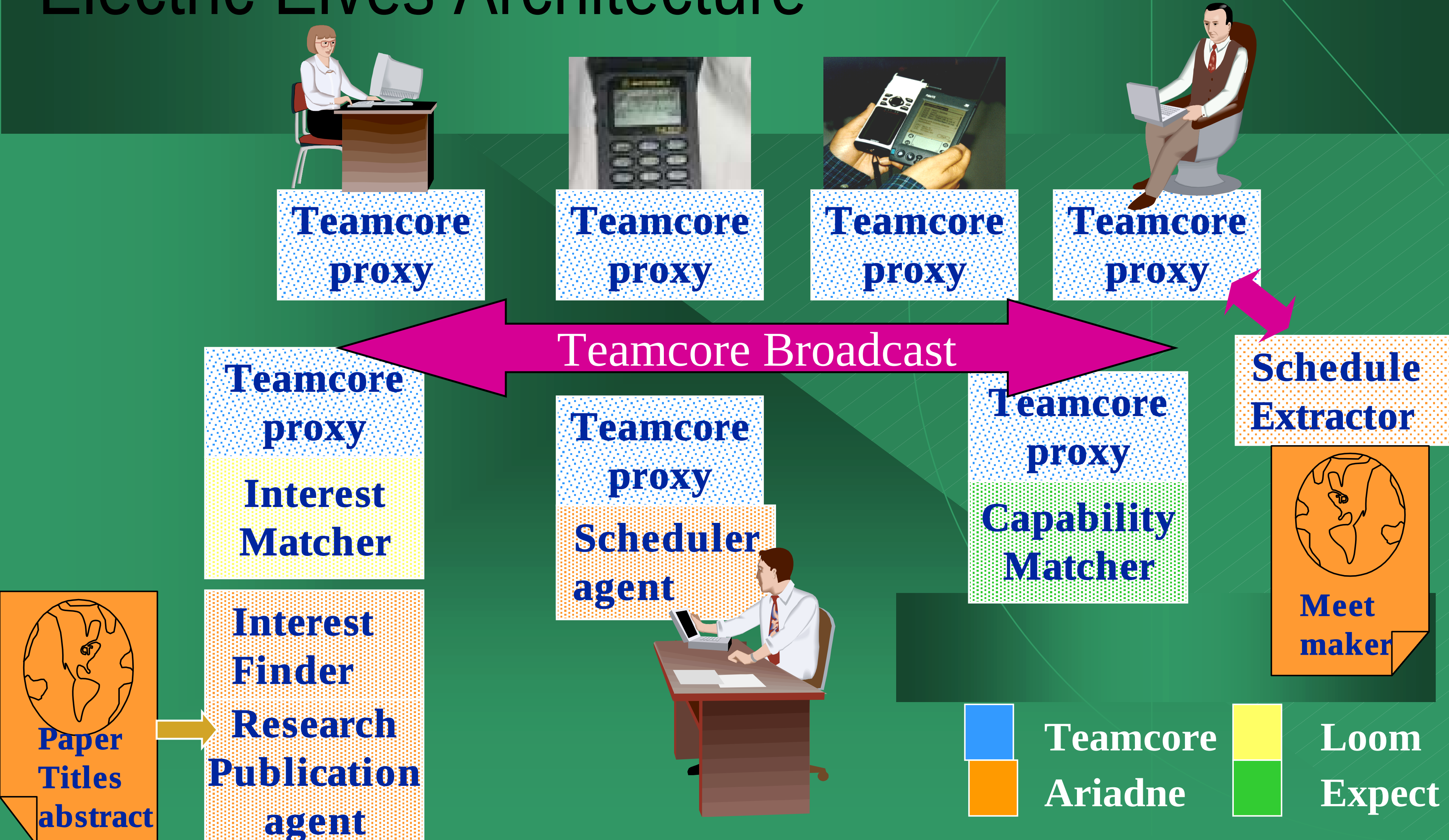
Outline

- Electric Elves
- MDPs for Adjustable Autonomy
- Safety Constraints for MDPs
- Results, Summary, Future Work

e-Elves

- Deployed MAS supporting collaboration at USC/ISI
- We want autonomous agents to:
  - perform tasks humans cannot do
  - automate tasks that humans can do
- Agent proxies helping users in daily activities:
  - location tracking
  - rescheduling meetings when delayed
  - assigning presenters for research meetings
  - ordering lunch

Electric Elves Architecture



Teamcore proxy

Teamcore proxy

Teamcore proxy

Teamcore proxy

Teamcore Broadcast

Schedule Extractor

Meet maker

Teamcore proxy

Interest Matcher

Teamcore proxy

Scheduler agent

Teamcore proxy

Capability Matcher

Paper Titles abstract

Interest Finder

Research Publication agent

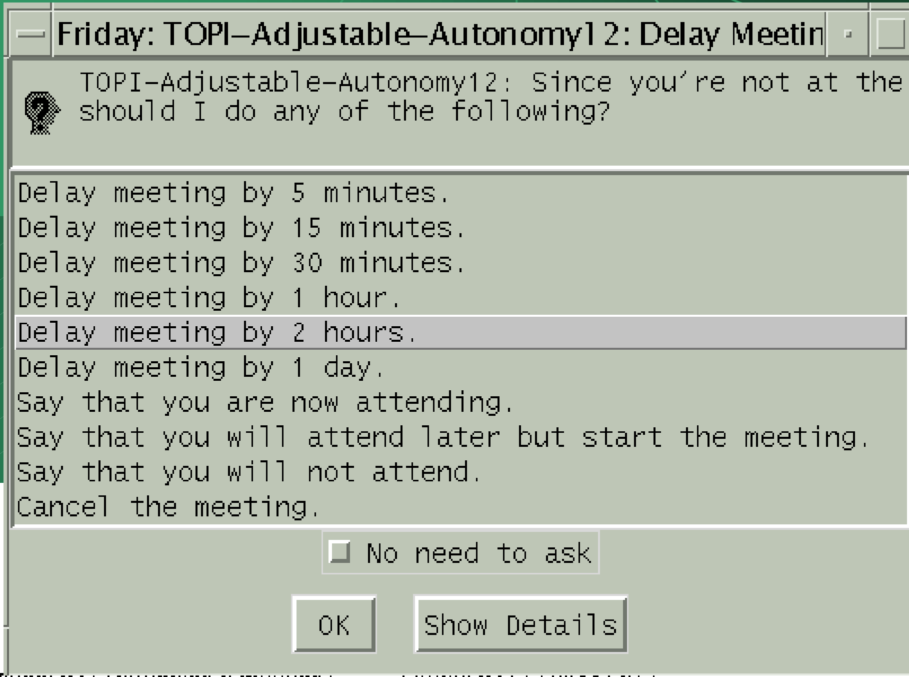
🟦 Teamcore

🟧 Ariadne

🟨 Loom

🟩 Expect

Monitoring Meetings



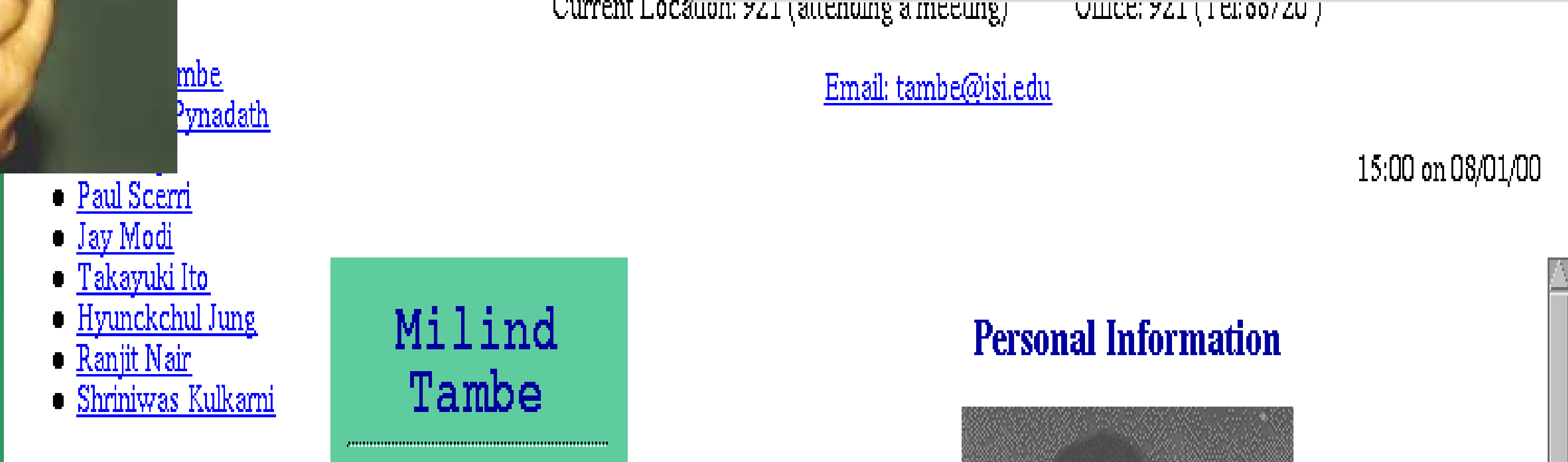
Friday: TOPI–Adjustable–Autonomy12: Delay Meetin

TOPI-Adjustable-Autonomy12: Since you're not at the
should I do any of the following?

Delay meeting by 5 minutes.
Delay meeting by 15 minutes.
Delay meeting by 30 minutes.
Delay meeting by 1 hour.
Delay meeting by 2 hours.
Delay meeting by 1 day.
Say that you are now attending.
Say that you will attend later but start the meeting.
Say that you will not attend.
Cancel the meeting.

□ No need to ask

le Locator



Current Location: 921 (attending a meeting)   Office: 921 (Tel:66720)

mbe
Pynadath

Email: tambe@isi.edu

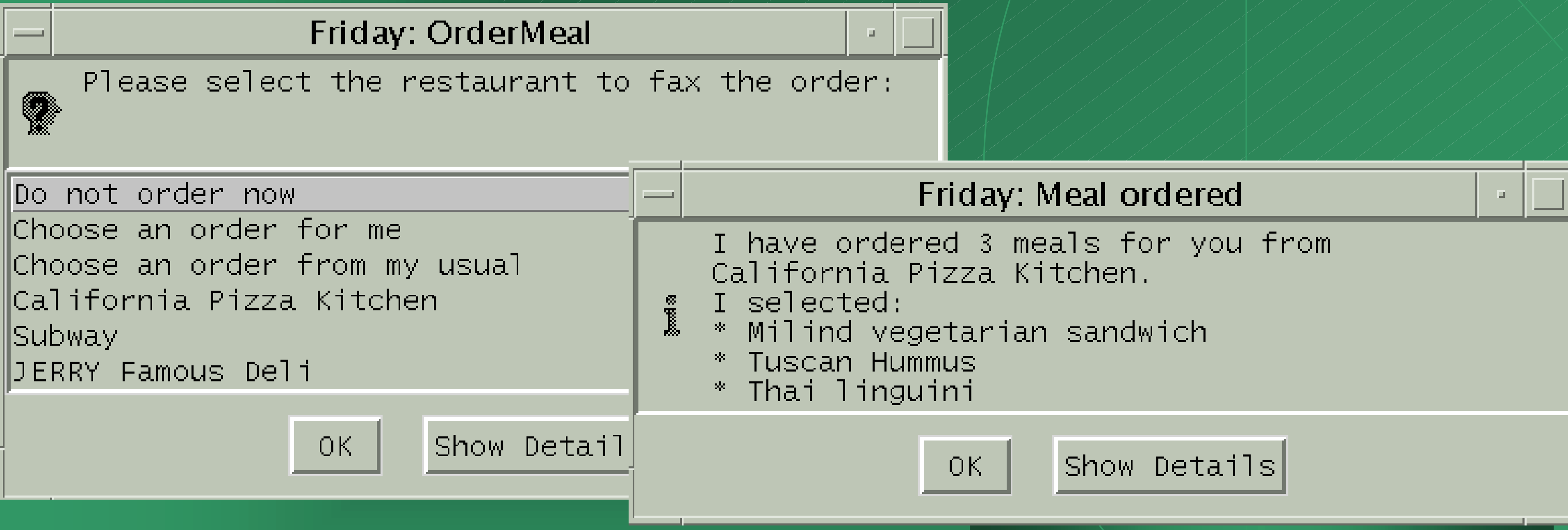
15:00 on 08/01/00

- Paul Scerri
- Jay Modi
- Takayuki Ito
- Hyunckchul Jung
- Ranjit Nair
- Shriniwas Kulkarni

Milind Tambe

Personal Information

Ordering Food



Friday: OrderMeal

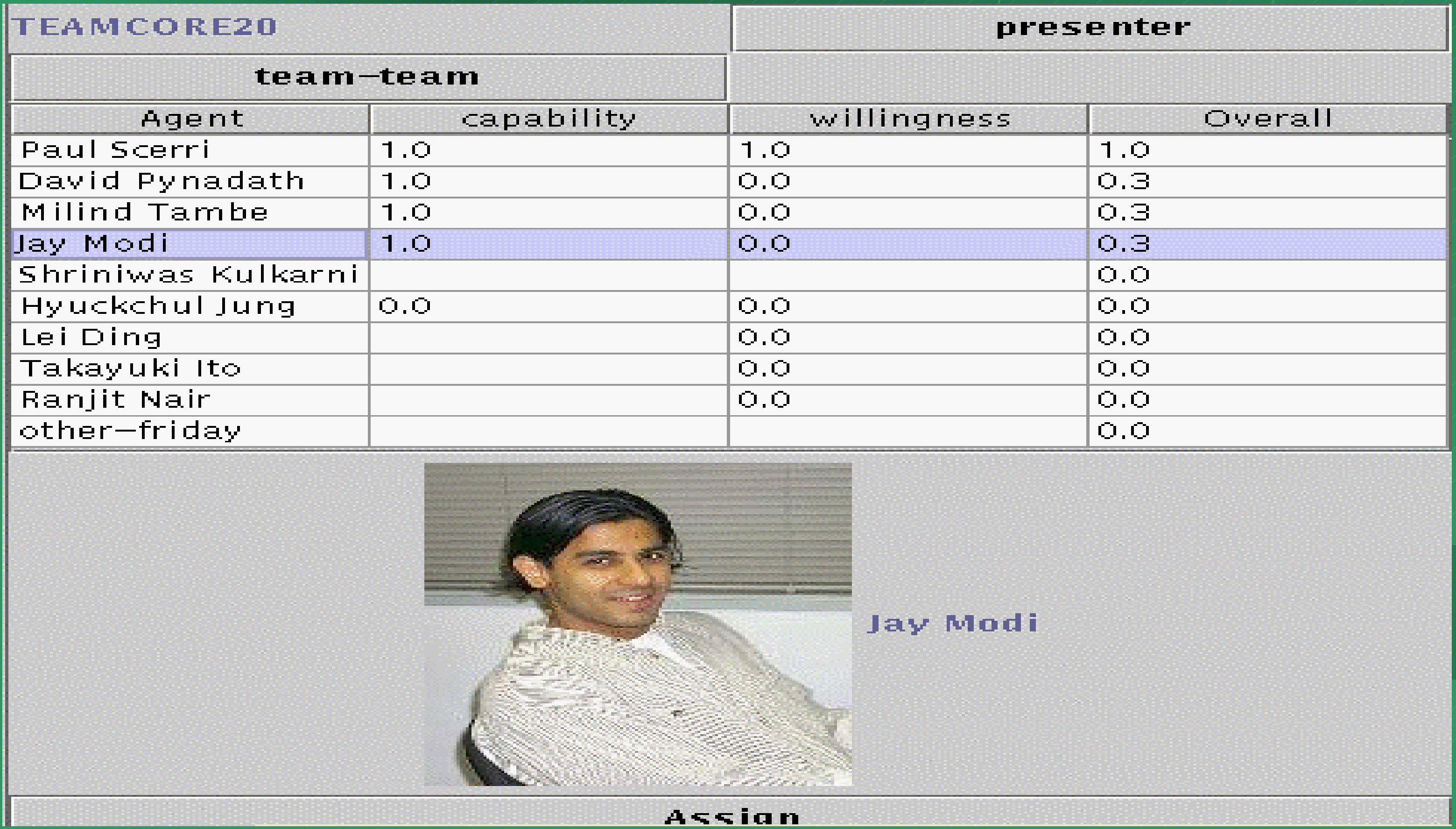
Please select the restaurant to fax the order:

Do not order now
Choose an order for me
Choose an order from my usual
California Pizza Kitchen
Subway
JERRY Famous Deli

Friday: Meal ordered

I have ordered 3 meals for you from California Pizza Kitchen.
I selected:
* Milind vegetarian sandwich
* Tuscan Hummus
* Thai linguini

Assigning Presenters



Agent	capability	willingness	Overall
Paul Scerri	1.0	1.0	1.0
David Pynadath	1.0	0.0	0.3
Milind Tambe	1.0	0.0	0.3
Jay Modi	1.0	0.0	0.3
Shriniwas Kulkarni			0.0
Hyuckchul Jung	0.0	0.0	0.0
Lei Ding		0.0	0.0
Takayuki Ito		0.0	0.0
Ranjit Nair		0.0	0.0
other-friday			0.0

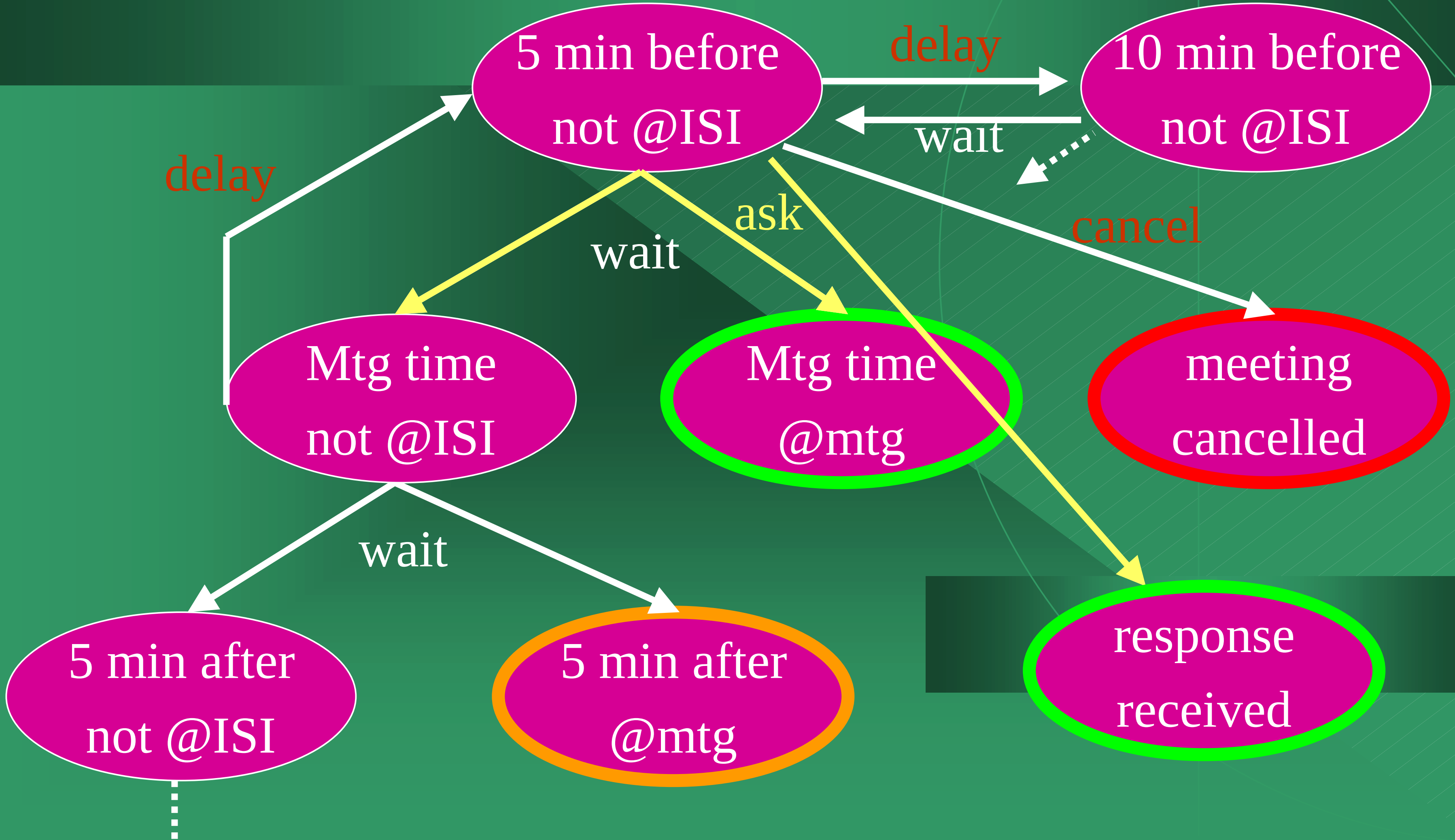
What is Adjustable Autonomy?

- Agents operating in a human organization:
  - act autonomously to save human effort
  - give up autonomy to avoid mistakes
- *Adjustable Autonomy (AA):*
  - “*Dynamically adjusting the level of autonomy of an agent depending on the situation*” [AAAI Spring Symp CFP 99]
- Key question:
  - When to transfer control/responsibility for decisions

Novel Issues in Team Settings

- Effects extend beyond individual user
  - Uncertainty in individual model
  - Actions that have global cost/benefit
  - ➔ Decision theory
- Flexibility in transfer of control: coordination challenge
  - User may not always be available to respond
  - Agent cannot wait indefinitely for response
  - ➔ Planning

Meeting Delay MDP



5 min before not @ISI

delay

10 min before not @ISI

wait

delay

ask

wait

cancel

Mtg time not @ISI

Mtg time @mtg

meeting cancelled

wait

response received

5 min after not @ISI

5 min after @mtg

Safety

- **Asimov's First Law of Robotics**: "A robot may not injure a human being, or, through inaction, allow a human being to come to harm."
- MDP Reward function can represent a notion of "safety", but...
  - No single reward function will satisfy all users
  - Learning personalized reward function may take a long time
- Instead, user provides agent with prior knowledge about safety
  - Must be easily expressed
  - Must have clear semantics

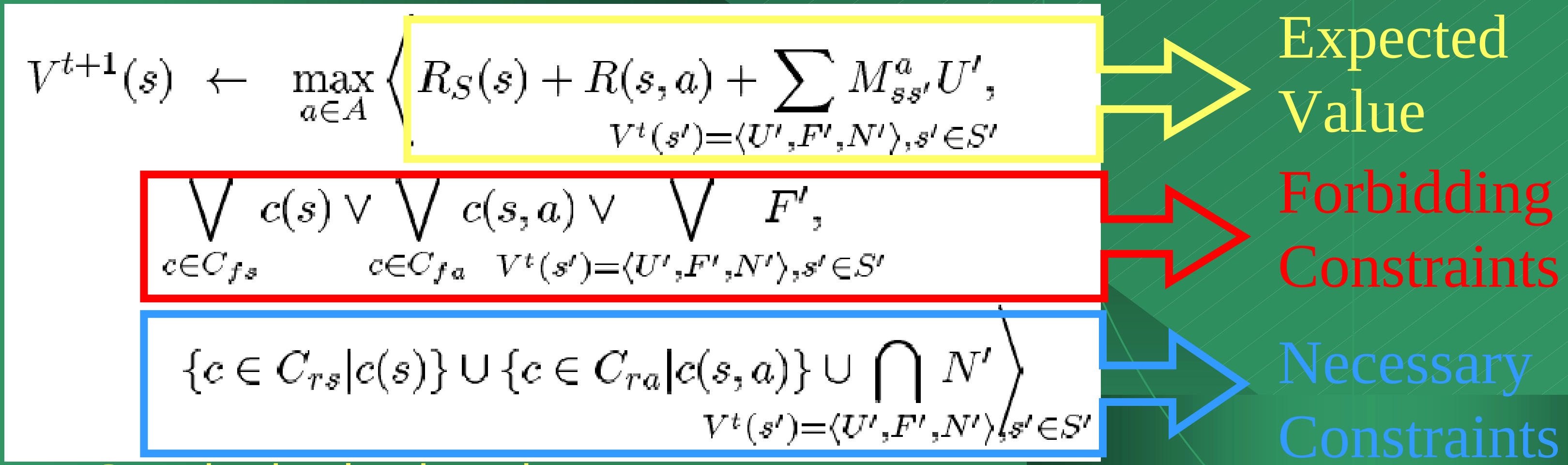
Constraints

- **Solution**: Individual users specify personalized ***constraints***,
  - User expresses strong preferences over actions and states
  - Analogous to Soar's prohibit and require preferences

	Forbidden (~)	Necessary (!)
States	It's past 3PM, but I have not eaten lunch	My teammates are informed of my status
Actions	Cancel meeting	Recharge battery

Constraint Propagation

- Value of state = <expected value, violation of forbidding constraints, satisfaction of necessary constraints>



$$V^{t+1}(s) \leftarrow \max_{a\in A}\Big\langle R_S(s)+R(s,a)+\sum_{V^t(s')=\langle U',F',N'\rangle,s'\in S'} M^a_{ss'}U',$$

$$\bigvee_{c\in C_{fs}} c(s) \vee \bigvee_{c\in C_{fa}} c(s,a) \vee \bigvee_{V^t(s')=\langle U',F',N'\rangle,s'\in S'} F',$$

$$\{c\in C_{rs}|c(s)\}\cup\{c\in C_{ra}|c(s,a)\}\cup \bigcap_{V^t(s')=\langle U',F',N'\rangle,s'\in S'} N'\Big\rangle$$

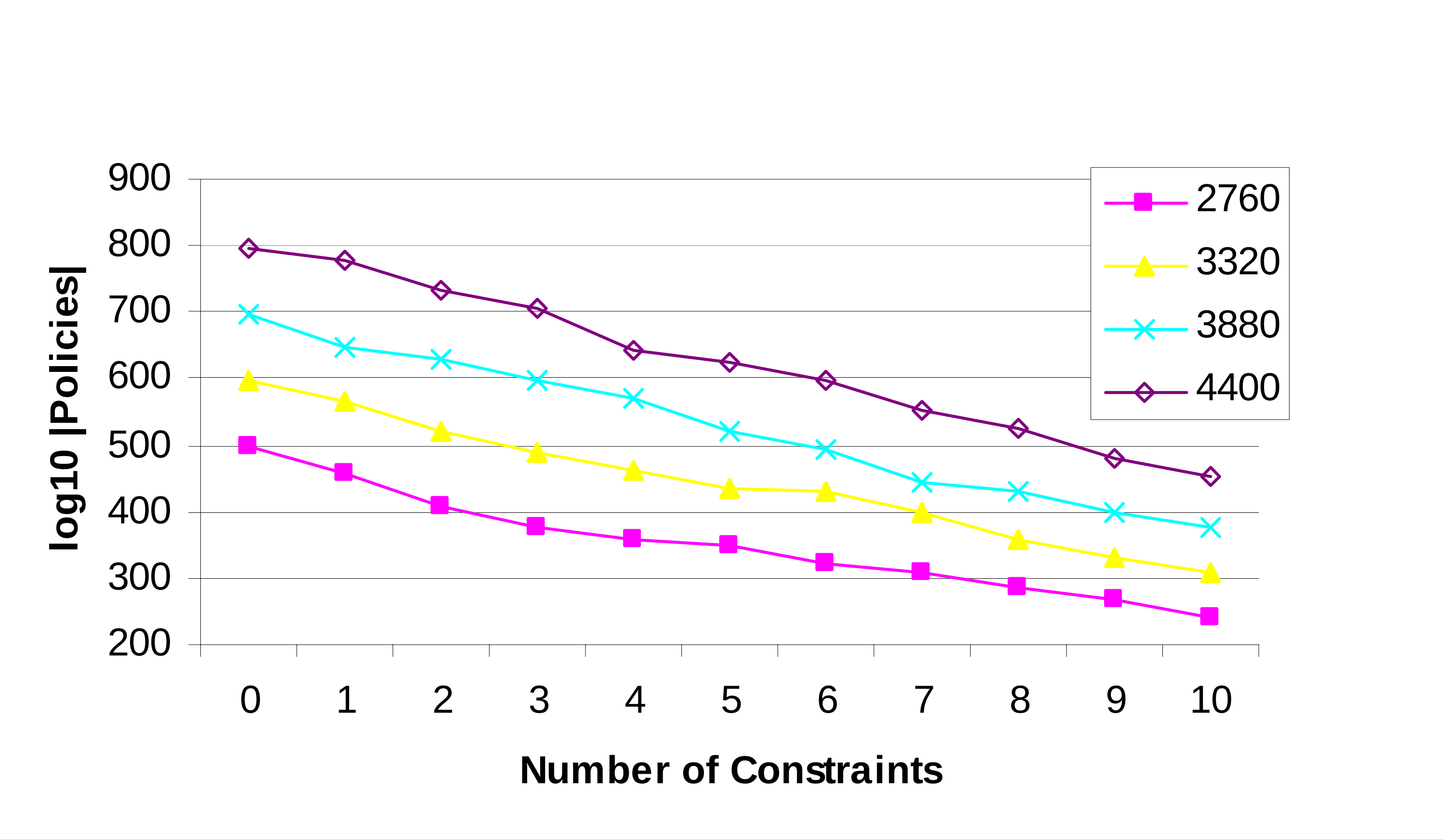
Expected Value

Forbidding Constraints

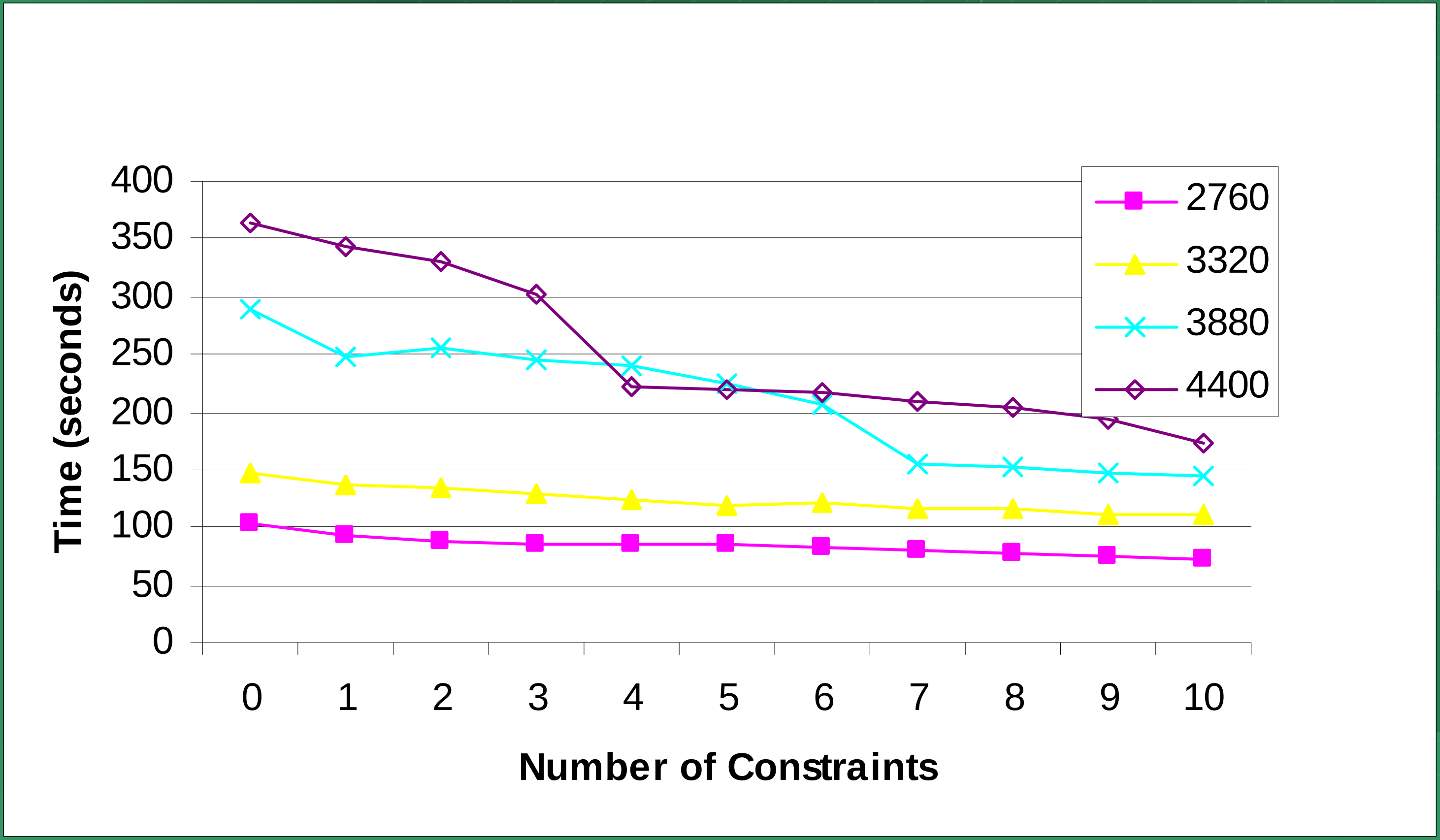
Necessary Constraints

- Standard value iteration
- Violated if state is forbidden or **ANY** child is forbidden
- Satisfied if state is necessary or **ALL** children are necessary

Elimination of Undesirable Behaviors



Policy Generation Time



Overall Electric Elves Results

- Multi-agent deployment in a real organization
  - Running 24/7 since June 1, 2000
  - No catastrophic failures
- Assists us in our daily activities
  - No emails about delays, cancels, etc.
  - No emails about scheduling talks at research meeting
  - Mobile devices extend interactions with agents
  - Fringe benefit: Friday is “active” reminder

Meetings Rescheduled

Meeting Resched	Unique meets	Person meets	Total resched	Auto resched	User resched
	387	642	346	208	138

Presenters Assigned

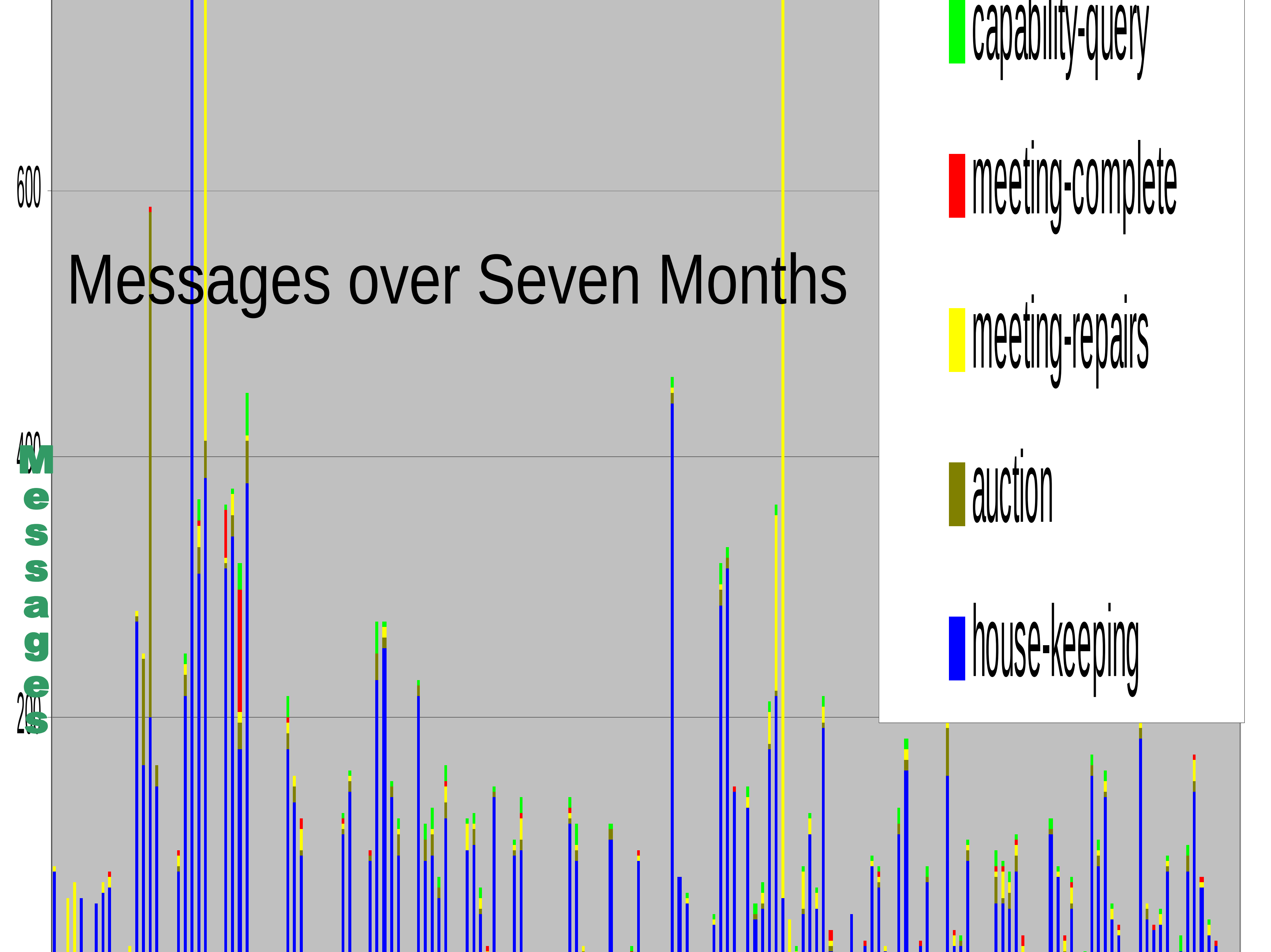
Presenter decisions	# meet	Auto decisions	Max bids	Avg bids
	10	8	9	6

Ongoing & Future Work

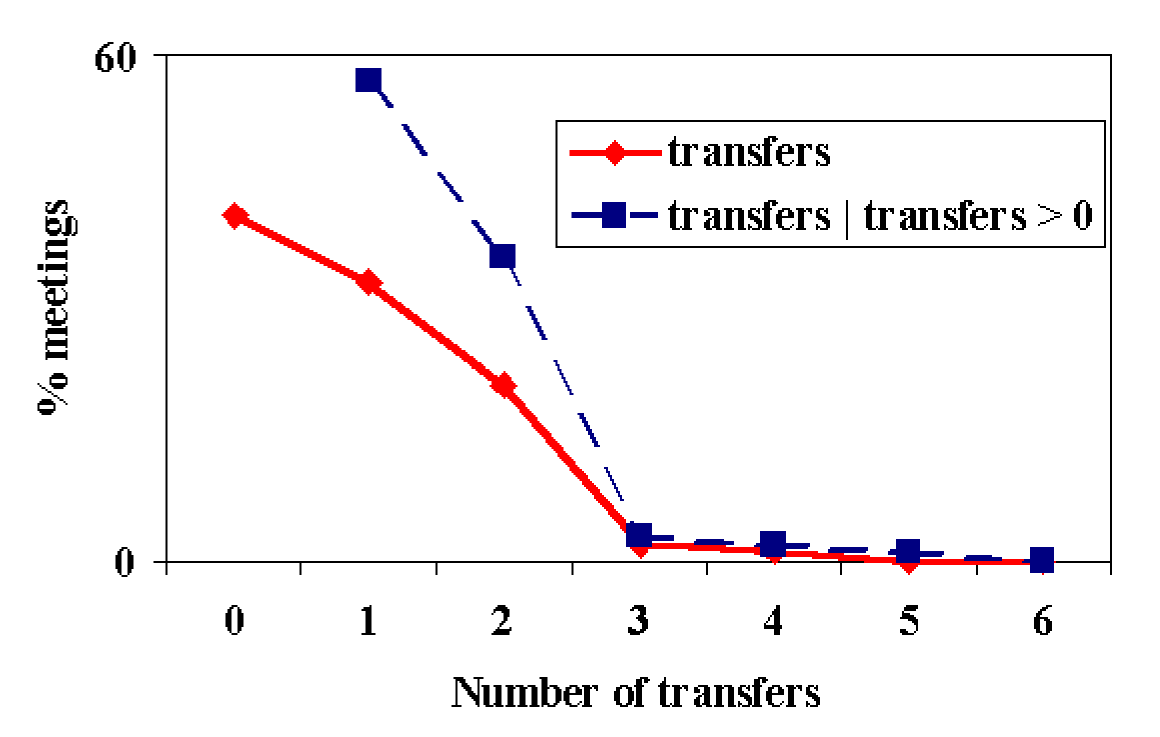
- Formalize general MDP model across decisions
- Evaluation of optimality of decisions

- Constraints that express other types of preferences
- Translate MDP policy into Soar rules

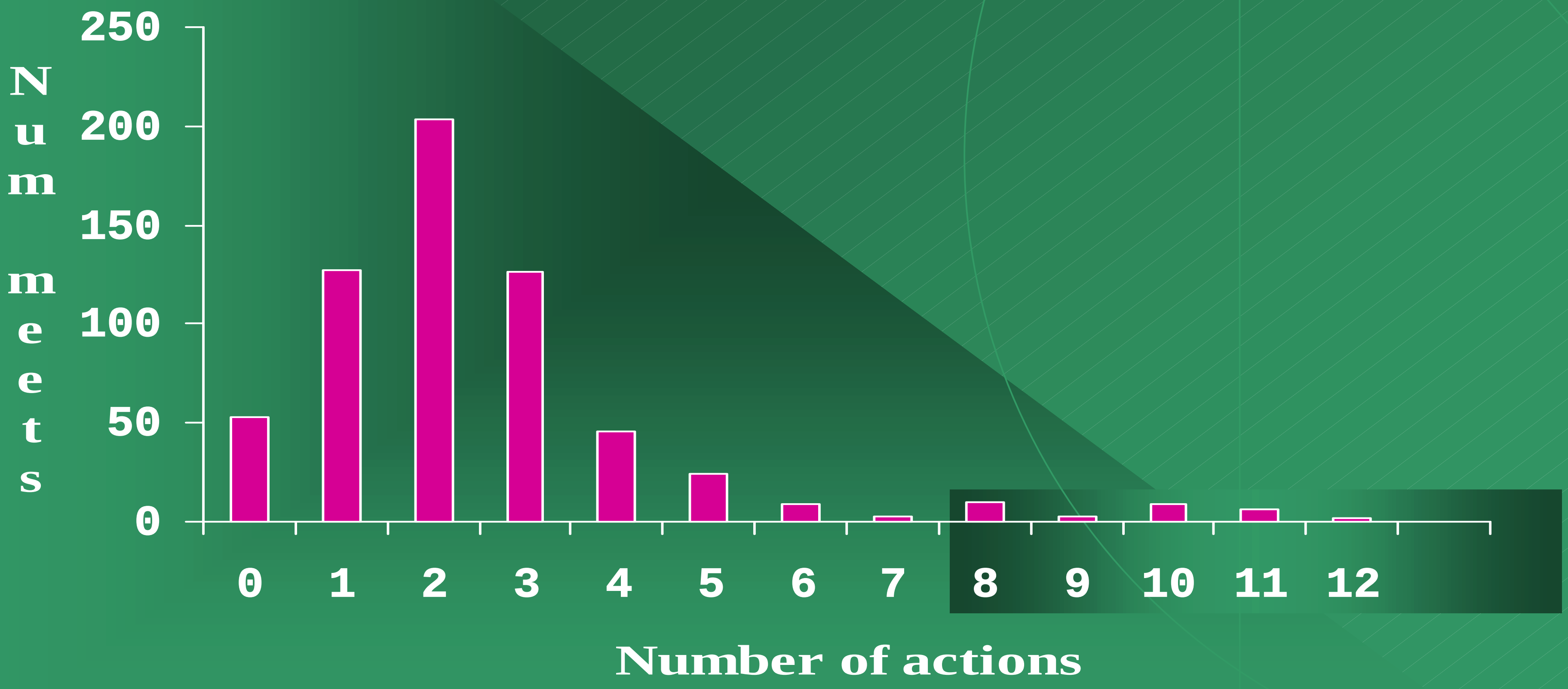
- http://www.isi.edu/agents-united

Messages over Seven Months
600
400
200
Messages
capability-query
meeting-complete
meeting-repairs
auction
house-keeping

Flexible Transfers of Control



Are multi-step policies actually used?


Num meets
Number of actions