Computational modeling of learning in complex problem solving tasks

Frédéric Dandurand

Presented at Seminars in Psychology
Department of Psychology, McGill University
October 10, 2007

Importance of problem solving

- Many activities of humans involve planning and problem solving
- “The ability to solve problems is one of the most important manifestations of human thinking.” (Holyoak, 1995)
- Classical research area in cognitive psychology

Information processing theory

- Describes problems in terms of states, transitions and constraints
  - Newell & Simon (1972)
  - Dominates problem solving
- Emphasizes search, heuristics and hints

Learning of problem solving tasks

- In cognitive psychology: how people learn or develop skills?

- Information processing theory has little to say about learning in problem solving
  - Imitation dismissed as rote memorizing (e.g., Katona, 1940)

Overview

- In this work on problem solving, learning is central
- Social learning
  - Imitation learning (learning by demonstration)
  - Verbal instructions (*if... then...* rules)
- Reinforcement learning (learning by trial and error)

Imitation learning

- Vast & active research area (human, animal, machine)
- Powerful & adaptive learning technique
- Cognitively complex
  - Understanding mentor's goals

    (Carpenter, Call, & Tomasello, 2002)
  - Complex, hierarchical problem representations

    (Byrne & Russon, 1998)
- Suggests understanding & incompatible with the "rote memorizing" view

Goals of experimental work

- Study learning in problem solving
  - Describe, quantify and compare learning under realistic learning regimes
  - Investigate cognitive mechanisms involved
- Study imitation learning in complex, planning-intensive tasks
  - Seek evidence that imitation is complex, and that "rote memorizing" view is inadequate
- Collect data to be modeled

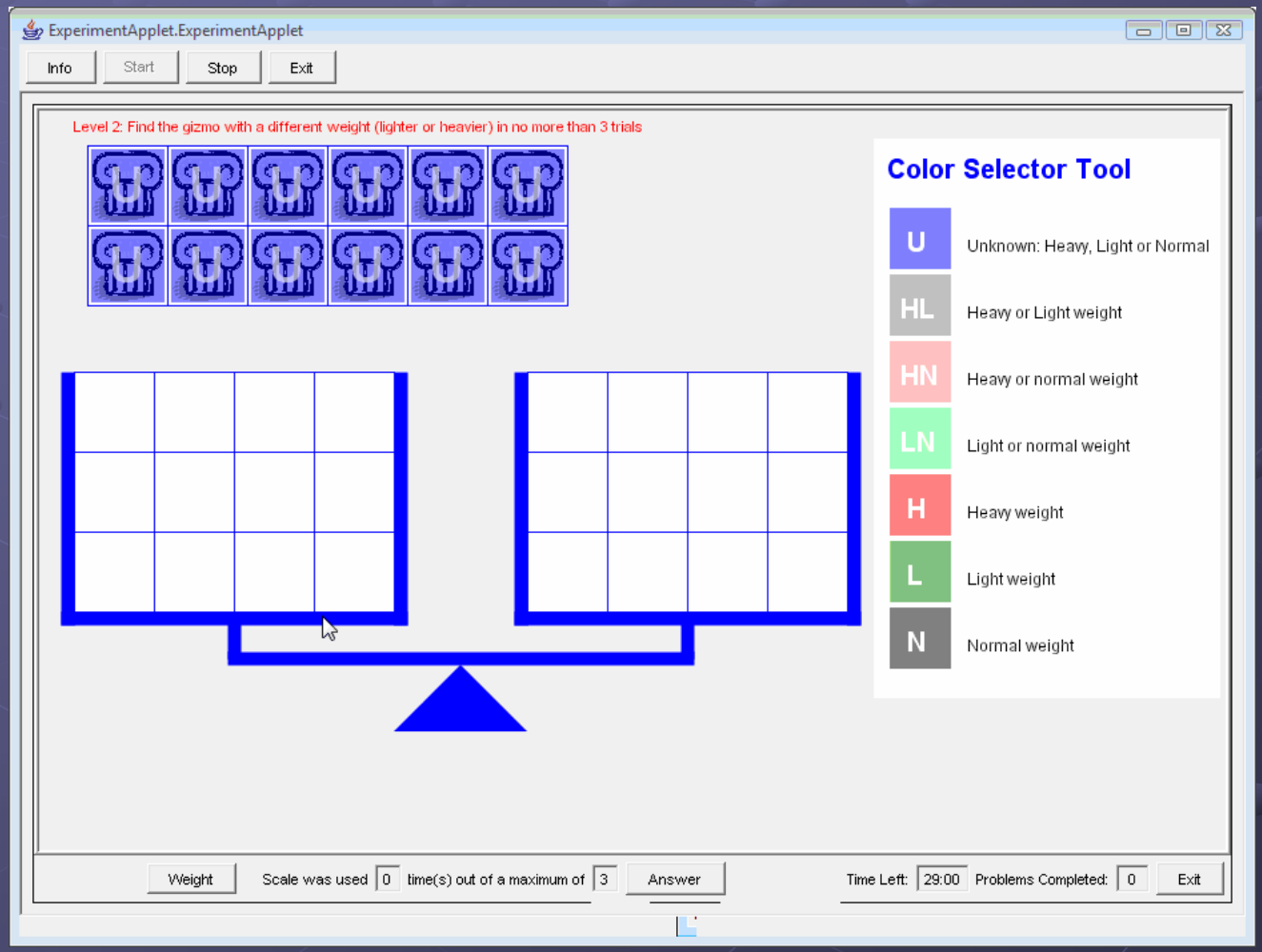
Goals of computational modeling work

- Investigate learning-centric models of problem solving
  - Coverage of human performance
  - Integration of multiple cognitive mechanisms
  - Cognitive plausibility
- Better understand mechanisms underlying problem solving

EXPERIMENTAL WORK

Gizmo problem solving task

Find, with **three uses of a scale**, the one gizmo that is either **heavier** or **lighter** than the rest of a set of 12 gizmos



Research questions

- How do people learn to solve problems under different learning regimes?
- How do participants evaluate the accuracy of their solutions?
- Is imitation about understanding or memorizing?

Research questions

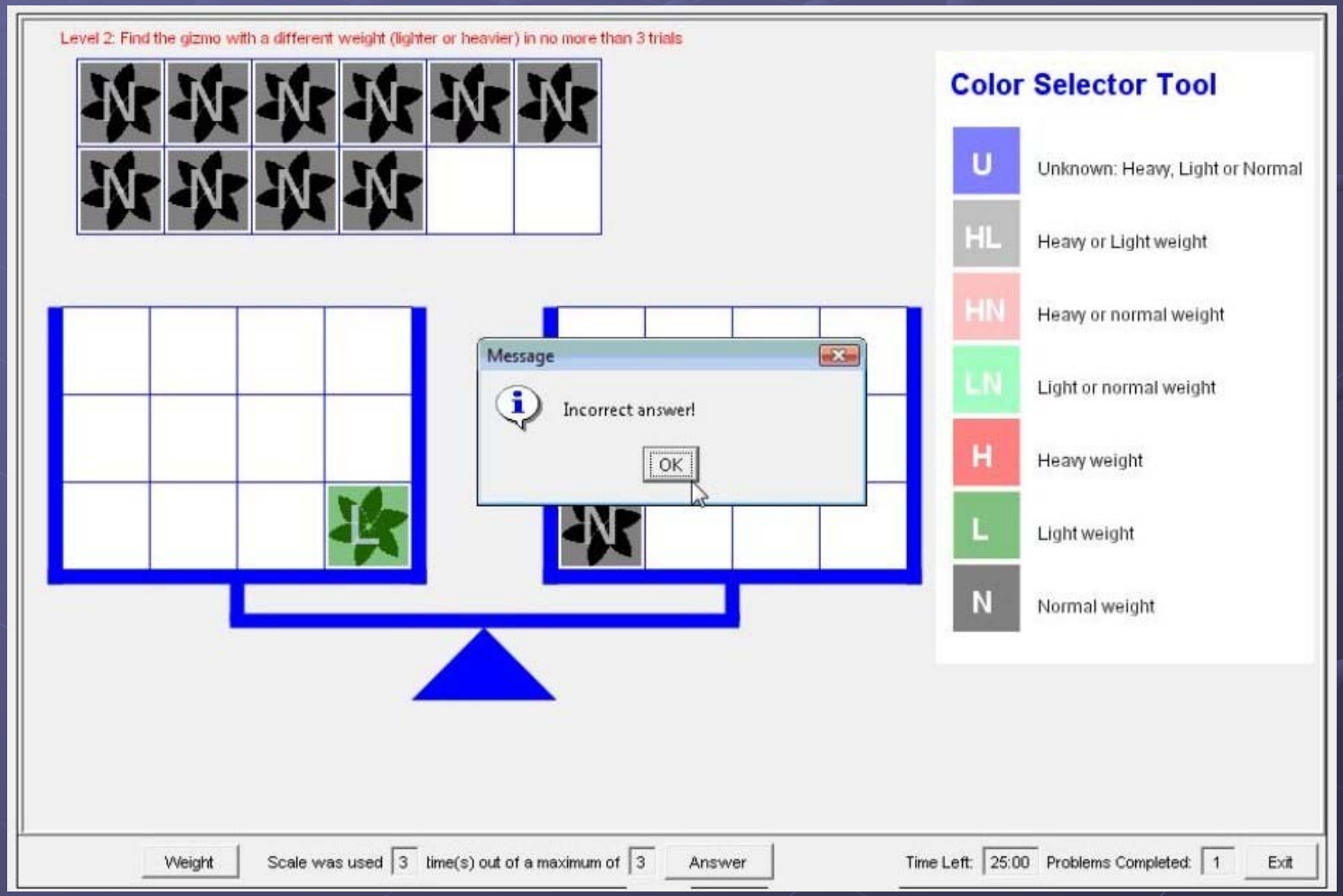
- How do people learn to solve problems under different learning regimes?
- How do participants evaluate the accuracy of their solutions?
- Is imitation about understanding or memorizing?

Experimental groups

- Reinforcement learning
  - Get binary rewards (answers correct or not)
- Imitation learning
  - Observe 5 demonstrations performed by expert
- Verbal instructions
  - Study instructions to solve the task (10 min.)

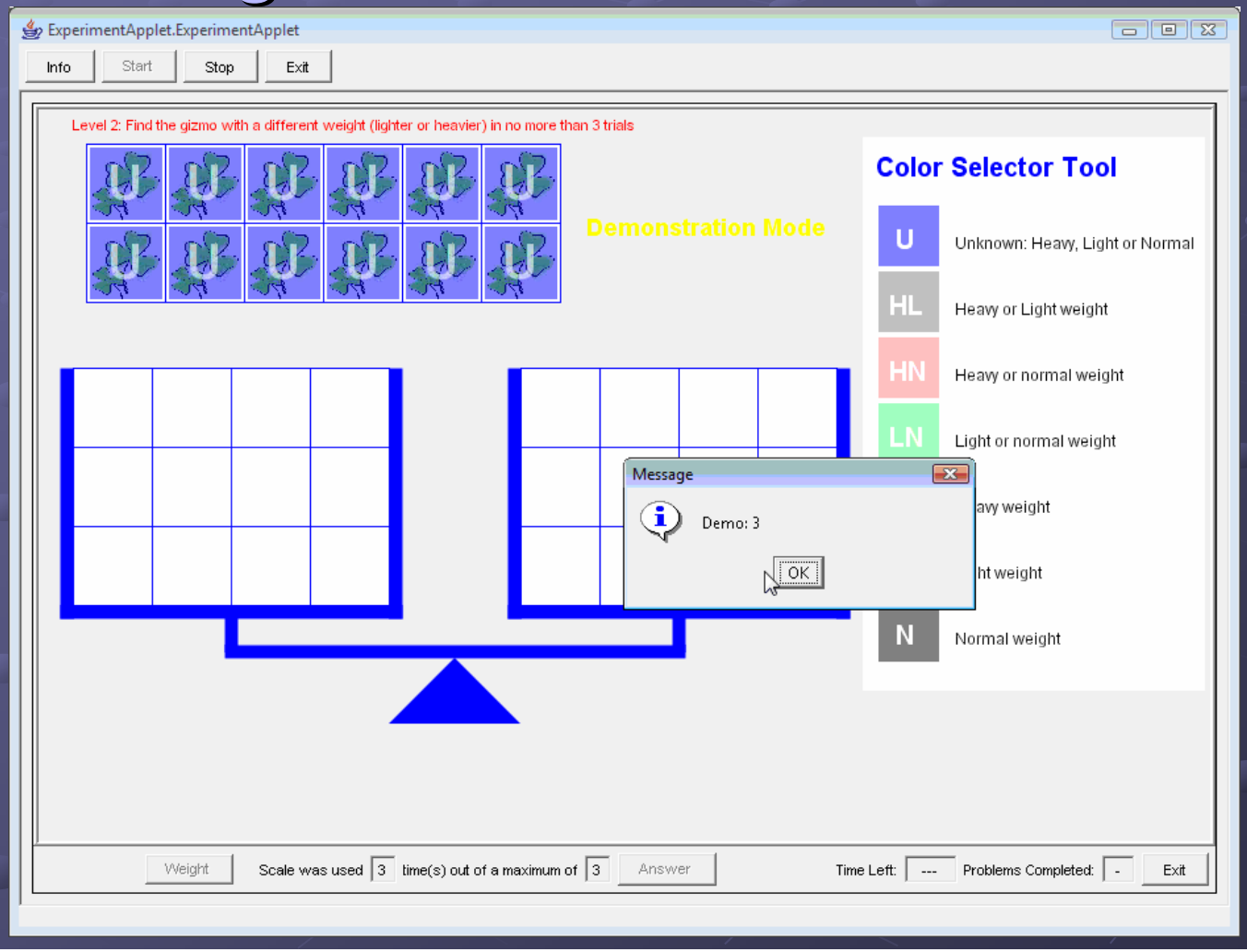
Reinforcement learning

Learning: told if answers are correct or not



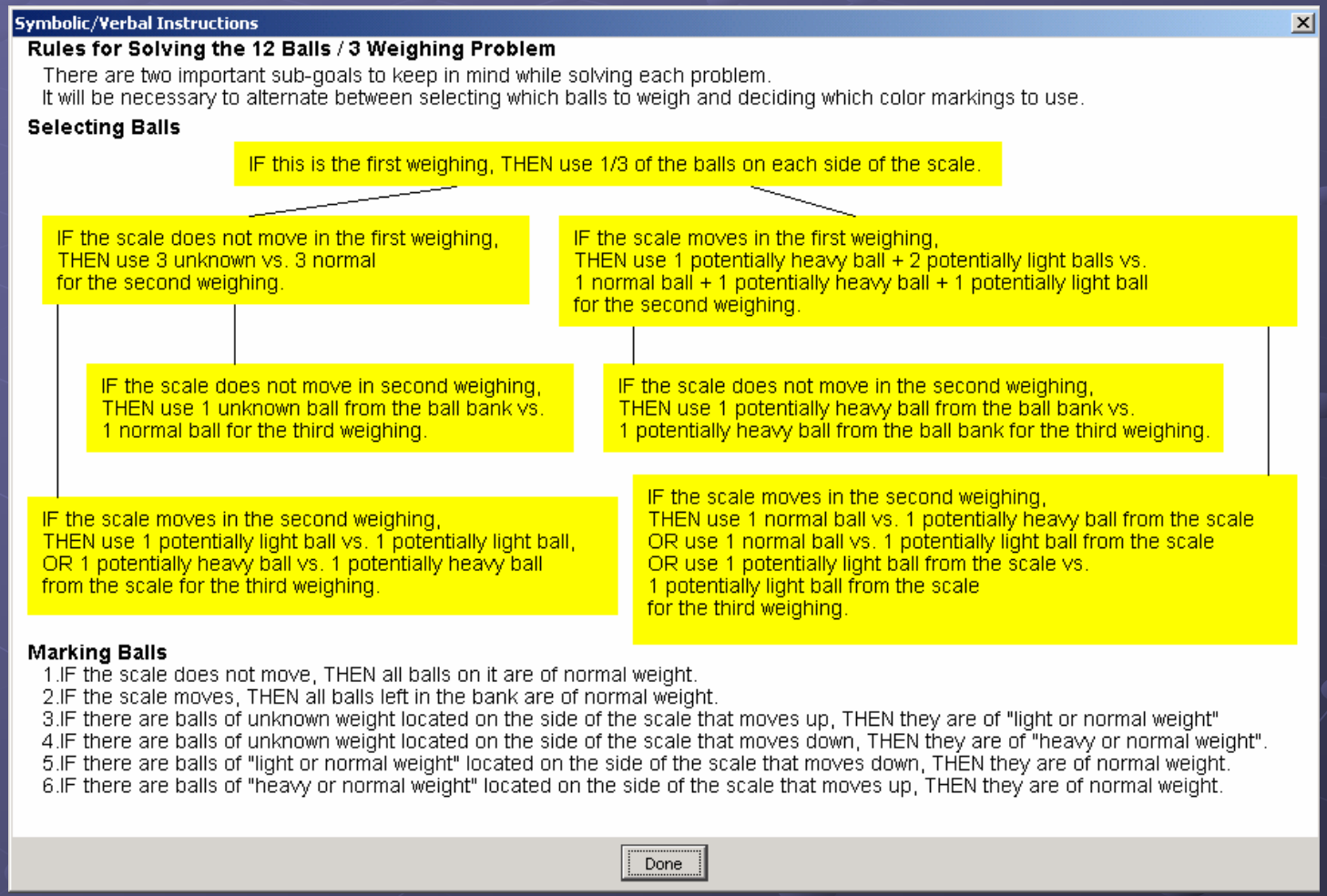
Imitation learning

Learning: shown five demonstrations

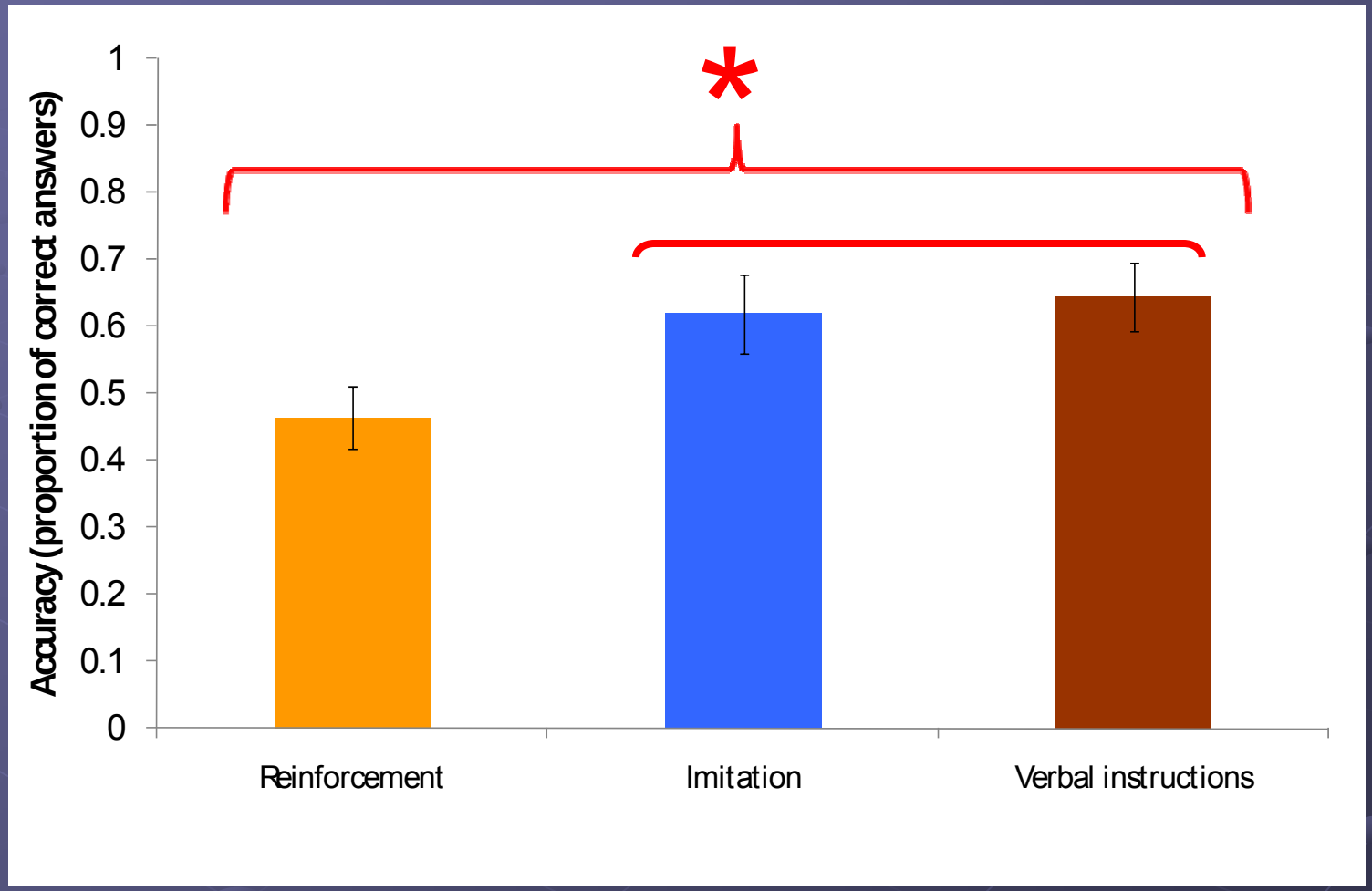


Verbal instructions

Learning: studied instructions



Learning under different training regimes



Imitation learning and verbal instruction groups were more accurate than the reinforcement learning group

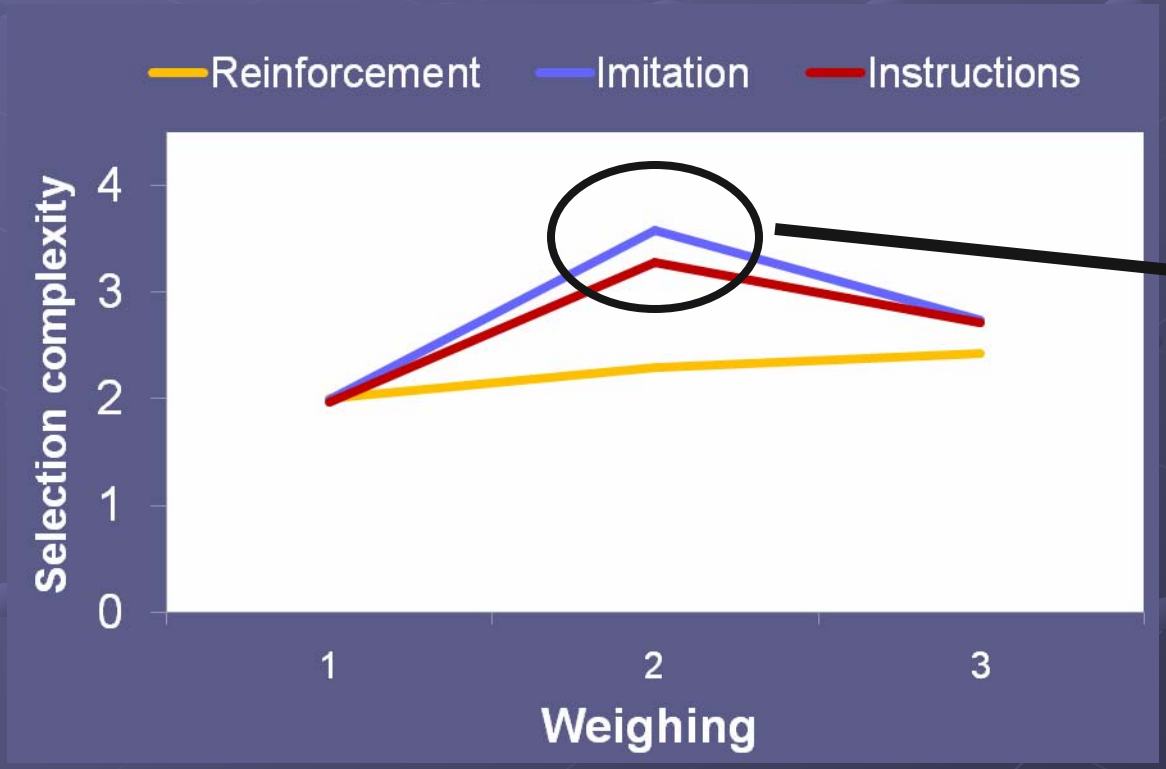
How do demonstrations and instructions help?

- General: Provide information -- show or describe how to solve problems

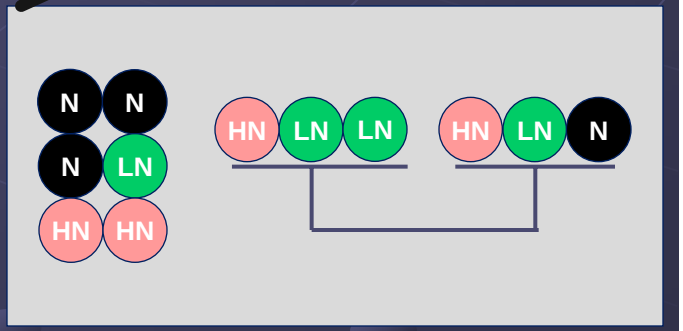
- More specifically: by possibly reducing...
  - Cognitive and perceptual biases (Freyd & Tversky, 1984)
  - Problem solving set (Glass & Holyoak, 1986)
  - Conceptual blocks (Adams, 1974)

Role of demonstrations and instructions

- Helped participants make complex selections
- Reduced simplicity bias



Difficult selection
(2nd weighing)



Research questions

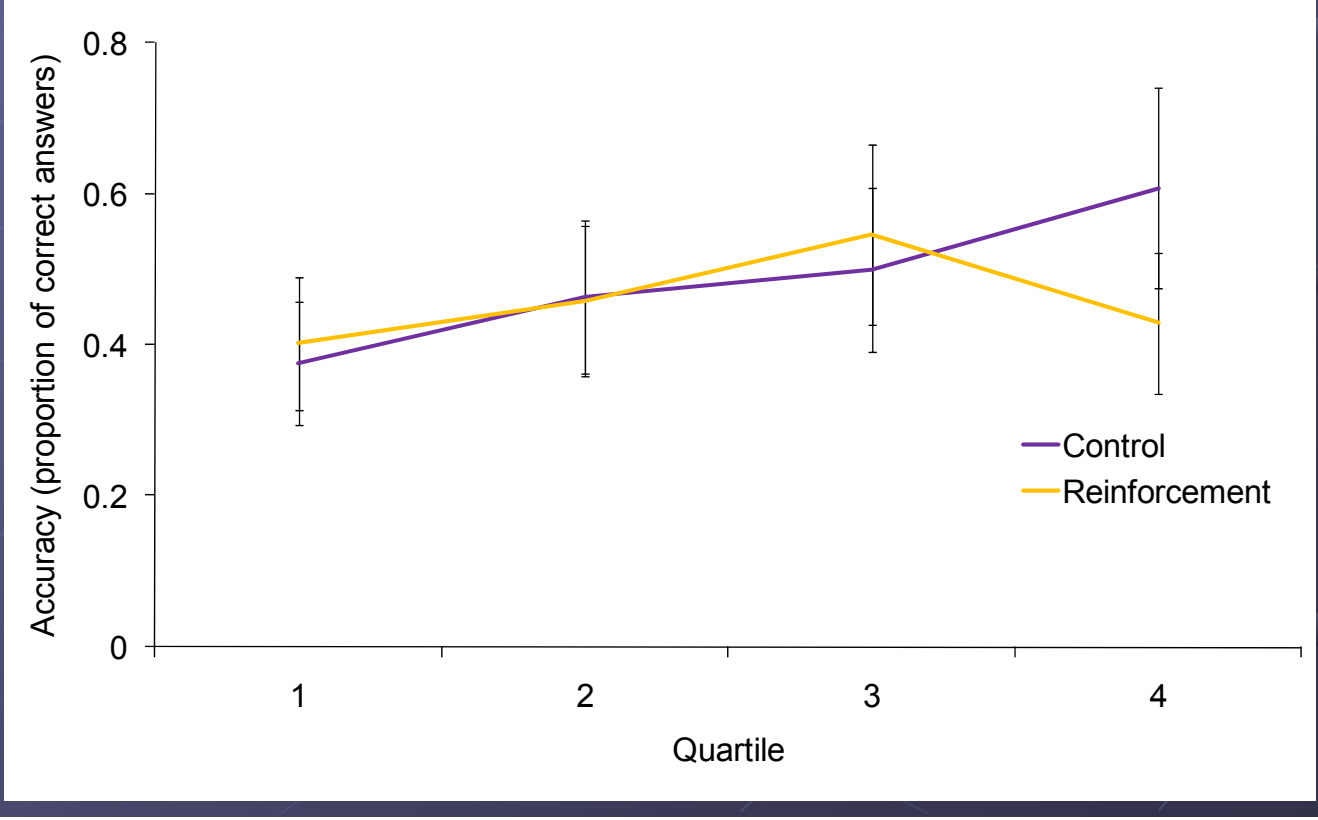
- How do people learn to solve problems under different learning regimes?
- How do participants evaluate the accuracy of their solutions?
- Is imitation about understanding or memorizing?

Using explicit feedback?

- Reinforcement learning group
  - Explicitly told if answers are correct or not

Is explicit feedback necessary?

- No – No significant difference between control and reinforcement
- Both groups improve with practice



How do participants evaluate the accuracy of their solutions?

- Reasoning & monitoring distance to goal
  - Used to self-evaluate solutions
  - Explicit rewards: Redundant

Research questions

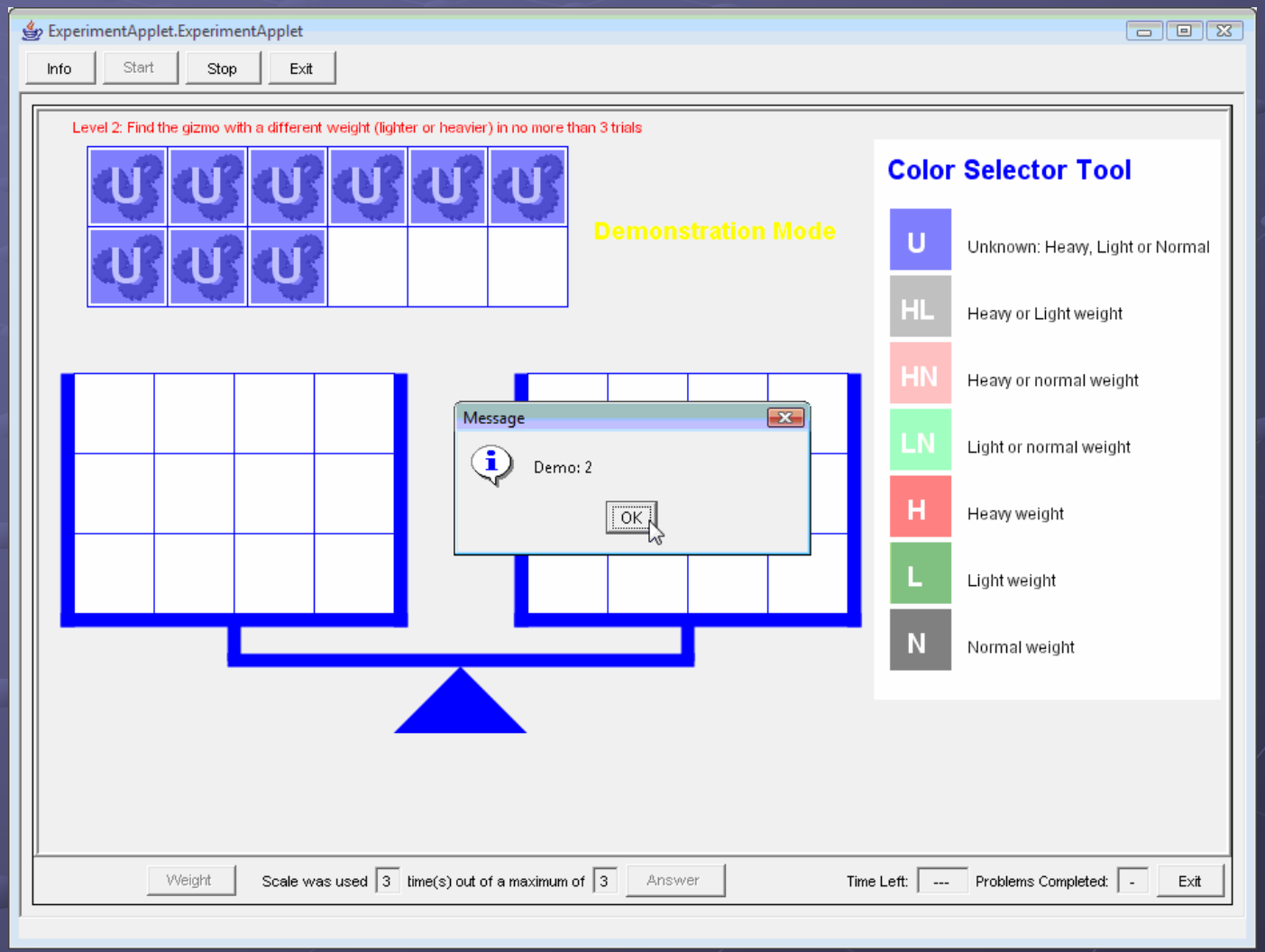
- How do people learn to solve problems under different learning regimes?
- How do participants evaluate the accuracy of their solutions?
- Is imitation about understanding or memorizing?

How to study understanding?

- Operational definition
  - Understanding = ability to generalize what was learned by observation to novel, more difficult problems
- Remove possibility of memorizing full solution

Generalization group

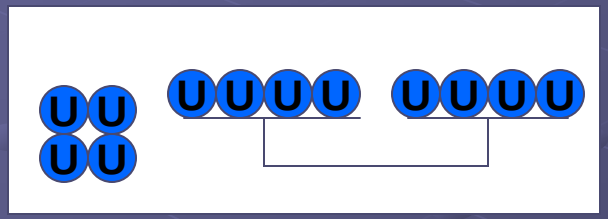
Learning: Watch 9 gizmo demos (solve 12 gizmo problems)



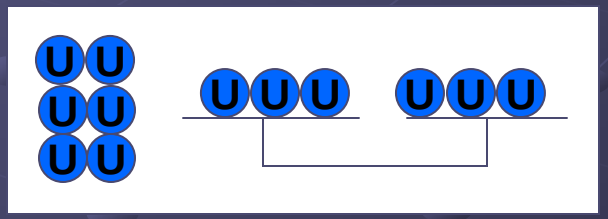
Strategies used on 1st weighing

- If participants memorize solutions, they should replicate what they observed in demonstrations

Imitation learning



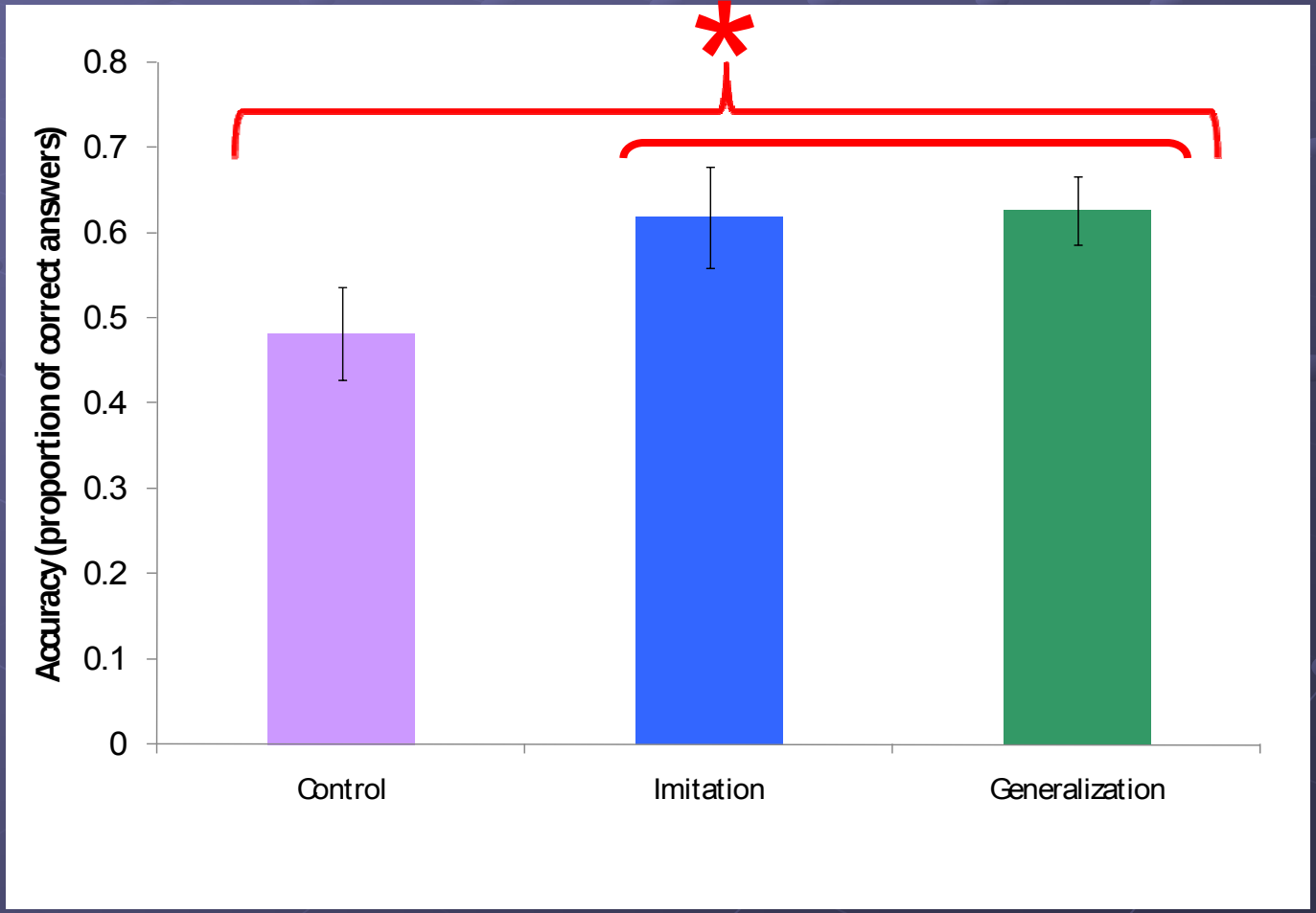
Generalization learning



- If select 3/3 in the 12 gizmo problems
  - then it is not always possible to solve without guessing

Generalization results

- Generalization group as accurate as imitation group



Strategies used on 1st weighing

	Experimental group		
Gizmos installed on each side			
1			
2			
3			
4			
5			
6			

Strategies used on 1st weighing

	Experimental group		
Gizmos installed on each side	**Control**		
1	**13.2%**		
2	**6.0%**		
3	**19.8%**		
4	**48.1%**		
5	**2.6%**		
6	**10.3%**		

- Control group: Variety of strategies

Strategies used on 1st weighing

	Experimental group		
Gizmos installed on each side	Control	**Imitation learning**	
1	13.2%	**0.0%**	
2	6.0%	**0.0%**	
3	19.8%	**0.0%**	
4	48.1%	**98.9%**	
5	2.6%	**1.1%**	
6	10.3%	**0.0%**	

- Imitation participants used correct (demonstrated) strategy

Strategies used on 1st weighing

	Experimental group		
Gizmos installed on each side	Control	Imitation learning	**Generalization**
1	13.2%	0.0%	**0.0%**
2	6.0%	0.0%	**0.9%**
3	19.8%	0.0%	**30.9%**
4	48.1%	98.9%	**65.4%**
5	2.6%	1.1%	**0.9%**
6	10.3%	0.0%	**1.9%**

- Almost 2/3 of generalization participants used to correct strategy

Take home messages

- Learning comparison
  - Imitation & verbal instructions groups more accurate than reinforcement
  - Demos & instructions reduced simplicity bias
- Evaluation of solution accuracy
  - No need for explicit feedback
  - Reasoning & monitoring distance to goal
- Correct generalization to more difficult task
  - Accuracy & use of correct strategy
  - Suggests understanding, not memorizing

COMPUTATIONAL MODELING WORK

Goals of modeling work

- Investigate learning-centric models of problem solving
  - Coverage of human performance
  - Integration of multiple cognitive mechanisms
  - Cognitive plausibility
- Better understand mechanisms underlying problem solving

Cognitive models of problem solving

- Current computational models: largely symbolic
  - General Problem Solver (Newell & Simon, 1963)
  - SOAR (Newell, 1990)
  - ACT-R (Anderson et al., 2004)
- New, learning-centric models
  - Connectionist, cascade-correlation (Shultz, 2003)
  - Reinforcement-learning based (Sutton & Barto, 1998)

Modeling principles

- Model the selection sub-task only
- Treat gizmos with same labels as a set (not distinguished)
- Describe problems in terms compatible with the information processing theory
  - States and actions (transitions)

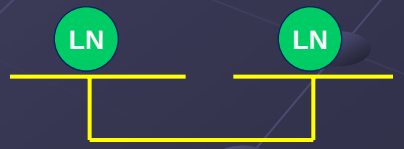
Problem description

- States: number of gizmos with each label
  - Example 1: 4N-4HN-4LN

  - Example 2: 12U

- Actions: what to install on each side of the scale (gizmo selection)
  - Example 1: 1HN-2LN/1HN-1LN-1N

  - Example 2: 1LN/1LN



How to model what humans do?

- Associate one action per state

State (if have this set of gizmos)	Action (what to put on scale)	
	Left	Right
4N-4HN-4LN →	1HN-2LN	1HN-1LN-1N
12U →	4U	4U

- For each state, evaluate possible actions & choose the best action

State (if have this set of gizmos)	possible actions		Use action with highest value given state
	Left	Right	
4N-4HN-4LN +	1HN-2LN	1HN-1LN-1N →	0.3
	1HN	1HN →	0.4
	2LN	2N →	0.1
	...	...	...

Two classes of models

- Action prediction (AP) models
  - Input: State
  - Output: Action (single, deterministic)
- Value prediction (VP) models
  - Input: State & Action
  - Output: Value of taking the action from state

Overview of models

Experimental group	Action prediction (AP) models	Value prediction (VP) models
Imitation learning	AP	VP-DPT (Direct Policy Training)
Reinforcement learning	N/A	VP-SARSA

Sibling Descendent Cascade Correlation

- Constructive neural networks (Fahlman & Lebiere, 1990)
  - Start small and grow as they learn
  - Recruit hidden units to increase computational power
  - Where recruits are installed?
    - Sibling: Deepest layer
    - Descendent: New layer

Models of imitation learning

Experimental group	Action prediction (AP) models	Value prediction (VP) models
Imitation learning	AP	VP-DPT (Direct Policy Training)
Reinforcement learning	N/A	VP-SARSA

- Use five demonstrations shown to humans as training data
- Compare
  - Accuracy (models and humans)
  - Performance
    - Training effort: how long to learn the task?
    - Model size: how many computational units recruited?

Action prediction model (AP)

- Sibling Descendent Cascade Correlation (SDCC) neural networks
  - Input: demonstrated state
  - Output: demonstrated action
- Training set: 15 patterns
  - 5 demonstrations x 3 weighings
- Task: For each state demonstrated, learn the demonstrated action

Value prediction model (VP-DPT)

- Sibling Descendent Cascade Correlation (SDCC) neural networks
  - Inputs: state & action
  - Output: value of taking action in state
  - Trained with Direct Policy Training (DPT)
- Training set: About 1600 patterns
  - Average of 106 possible actions per state visited

Value prediction model (VP-DPT)

- Task: For each state demonstrated, learn that the demonstrated action is the best possible one
- Direct Policy Training (DPT) algorithm
  - Initialize networks with random connection weights
  - While demonstrated action does not rank first
    - Increase value of demo action by some learning rate
    - Decrease values of non-demo actions ranking higher than demo

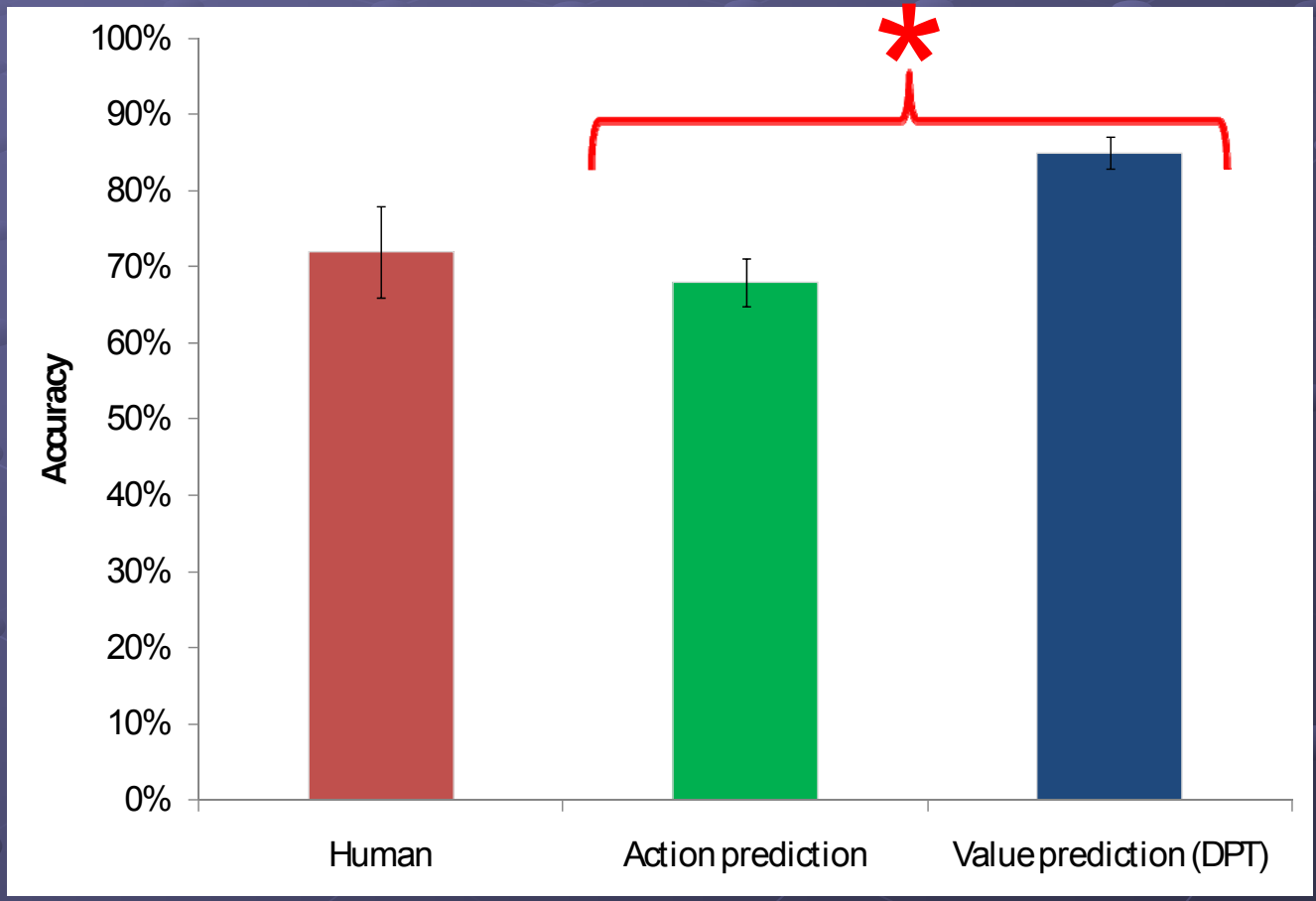
State (if have this set of gizmos)	Possible actions		Action value	
	Left	Right	Init	Final
4N-4HN- 4LN	1HN-2LN	1HN-1LN-1N	-1.432	0.5
	1HN	1HN	0.023	-0.02
	2LN	2N	0.312	0.12
	...	...	...	...

Training and testing

- Training
  - 5 demonstrations
    - Matches imitation condition
  - 24 demonstrations
    - Networks get complete information (indeed, they reach 100% accuracy)
    - How models learn task completely from demos
- Testing
  - 24 problems
- Accuracy: percentage of correct answers (out of 24)

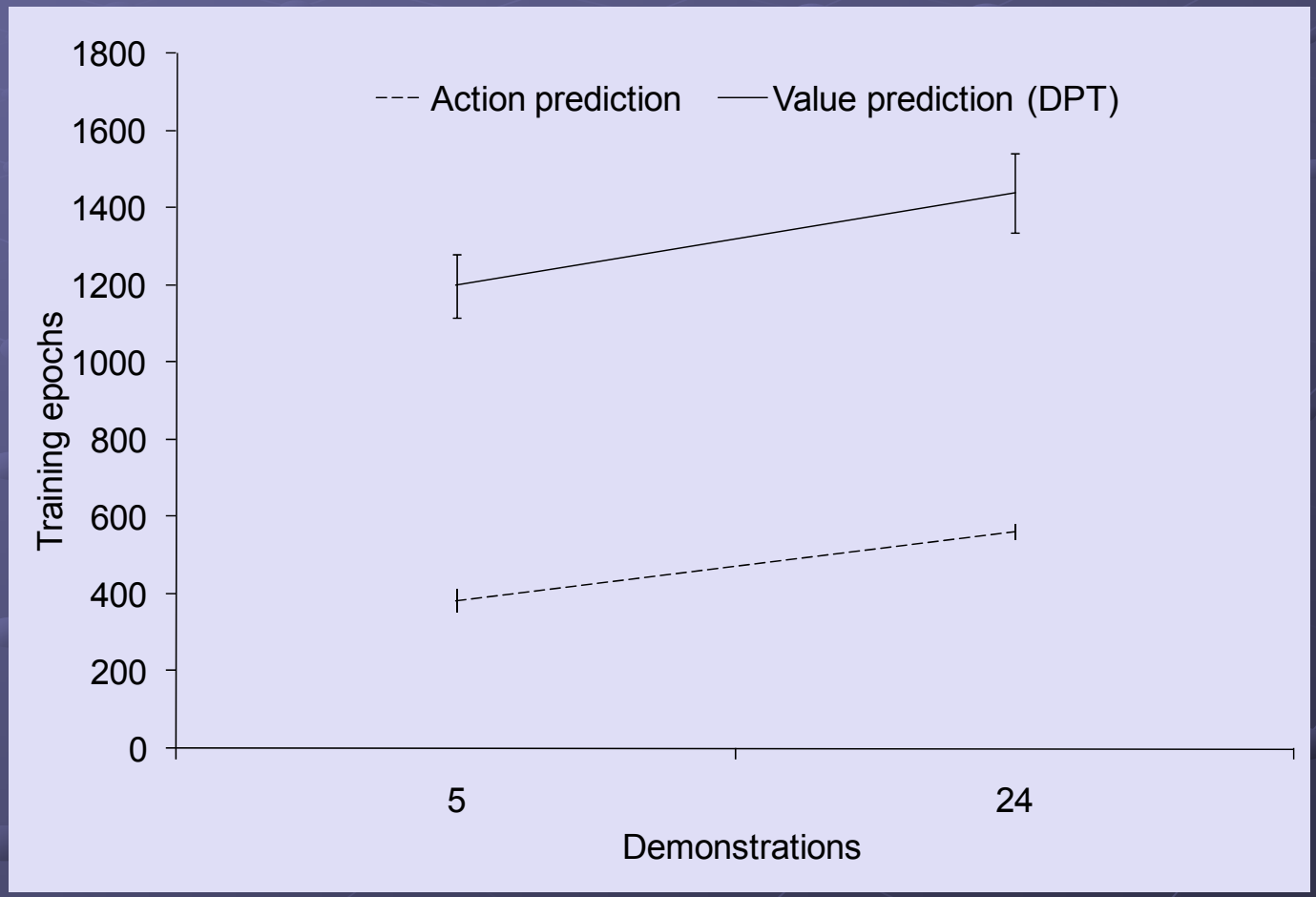
Results: Imitation accuracy

- Neither model differs from human accuracy
- Value prediction models > Action prediction



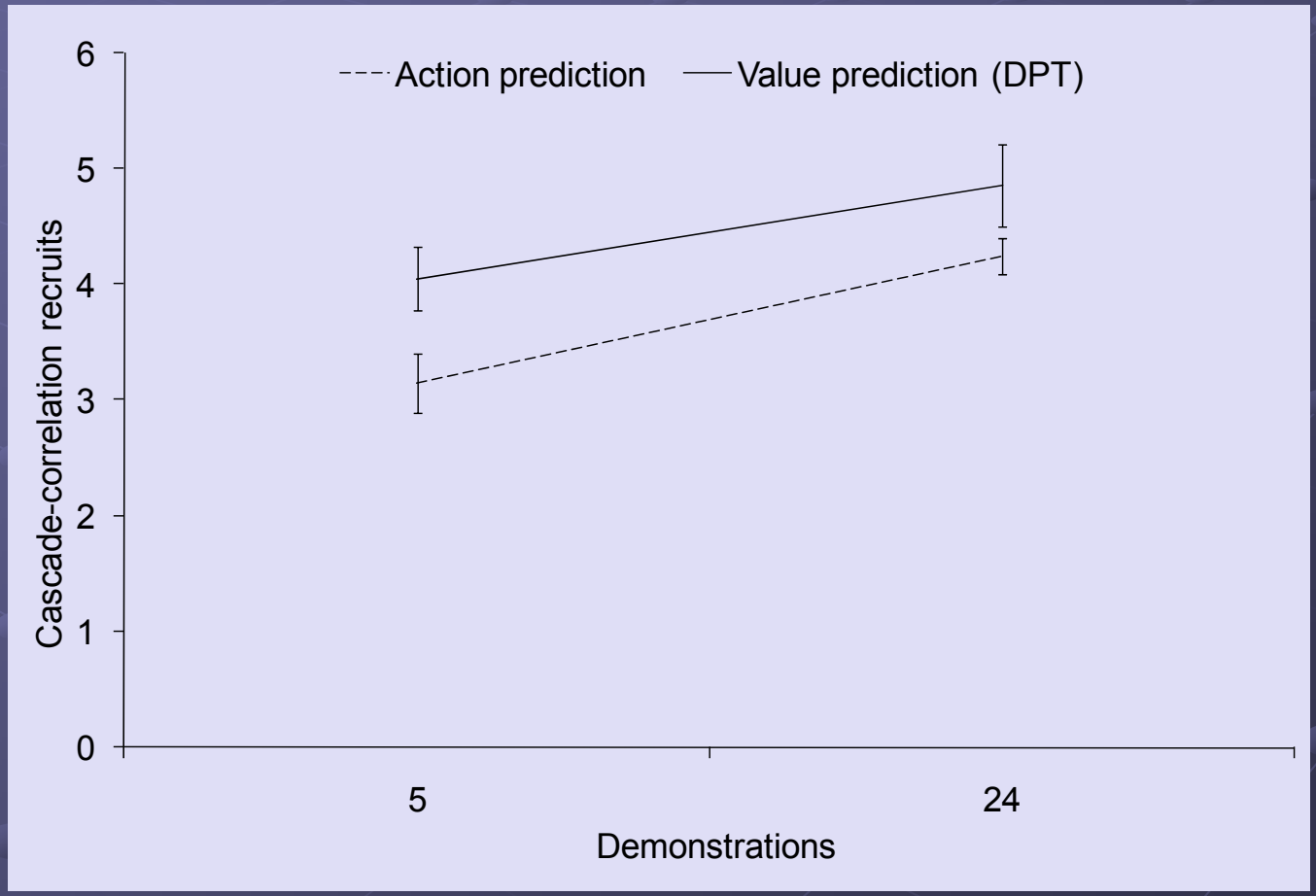
Results: Training effort (epochs)

- Action prediction models < Value prediction
- 5 demonstrations < 24 demonstrations



Results: Model size (recruits)

- Action prediction models < Value prediction
- 5 demonstrations < 24 demonstrations



Imitation model summary

Criterion	Action prediction	Value prediction
Match human accuracy	+	+

Imitation model summary

Criterion	Action prediction	Value prediction
Match human accuracy	+	+
More accurate		+

Imitation model summary

Criterion	Action prediction	Value prediction
Match human accuracy	+	+
More accurate		+
Trains faster	+	

Imitation model summary

Criterion	Action prediction	Value prediction
Match human accuracy	+	+
More accurate		+
Trains faster	+	
More compact models (fewer recruits)	+	

Models of reinforcement learning

Experimental group	Action prediction models	Value prediction models
Imitation learning	AP	VP-DPT (Direct Policy Training)
Reinforcement learning	N/A	VP-SARSA

- Use environmental rewards (answer correct or not) to train model

Task characteristics

- Learning by reinforcement is much more difficult than learning by demonstration
- Why?
  - Impoverished information

Information	Imitation task	Reinforcement task
Rate (frequency)	After each weighing	After 3rd weighing only *If solution fails, unclear which action was poor*
Information content	What selection action to take	Binary evaluation signal *If solution fails, not told what selection action should be taken instead*

Value prediction model overview

- Similarities with VP-DPT model
  - Input: State & action; Output: expected value
  - Sibling Descendent Cascade Correlation
- Differences: More complex task & model
  - Sequence of actions (1st, 2nd and 3rd weighings) lead to success or failure, but no info on individual actions
  - Additional system to propagate environmental rewards for 1st and 2nd weighings

Propagating rewards

- SARSA $(s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1})$ (Sutton & Barto, 1998)

  $Q(s_t,a_t) \leftarrow Q(s_t,a_t)+\alpha[r_{t+1}+\gamma Q(s_{t+1},a_{t+1})-Q(s_t,a_t)]$

  - Temporal difference (TD) learning algorithm
  - Learns to accurately predict expected rewards (Quality, Q) using
    - Reward actually obtained in the next state
    - Discrepancy between estimate of values of current state and next state
- Biological plausibility: SARSA-like mechanisms found in the brain

  (Samejima, Ueda, Doya, & Kimura, 2005; Suri & Schultz, 1999; Houk, Adams, & Barto, 1995)

Two levels of learning

- SARSA: estimates reward (Q value)
  - Initially poor
  - Improves over time (with exploration)
- SDCC function approximator learns to approximate current reward estimates
  - Subordinate system to SARSA

Action selection method

- Value prediction models can select actions flexibly
  - Not necessarily the best possible action
- Used a modified Softmax method
  - Softmax: action has higher expected value → higher probability of taking that action
  - Keep only the *n* best actions (with highest predicted rewards) from a given state
- ≈ Working memory
  - Number of active elements

Working memory

- Varied number of actions used in Softmax (≈ working memory size)
  - n=1 (Hardmax), 3, 5 and 10
- Estimates of human working memory capacity
  - 7 +/- 2 (Miller, 1956)
  - 4 +/- 1 (Cowan, 2000)

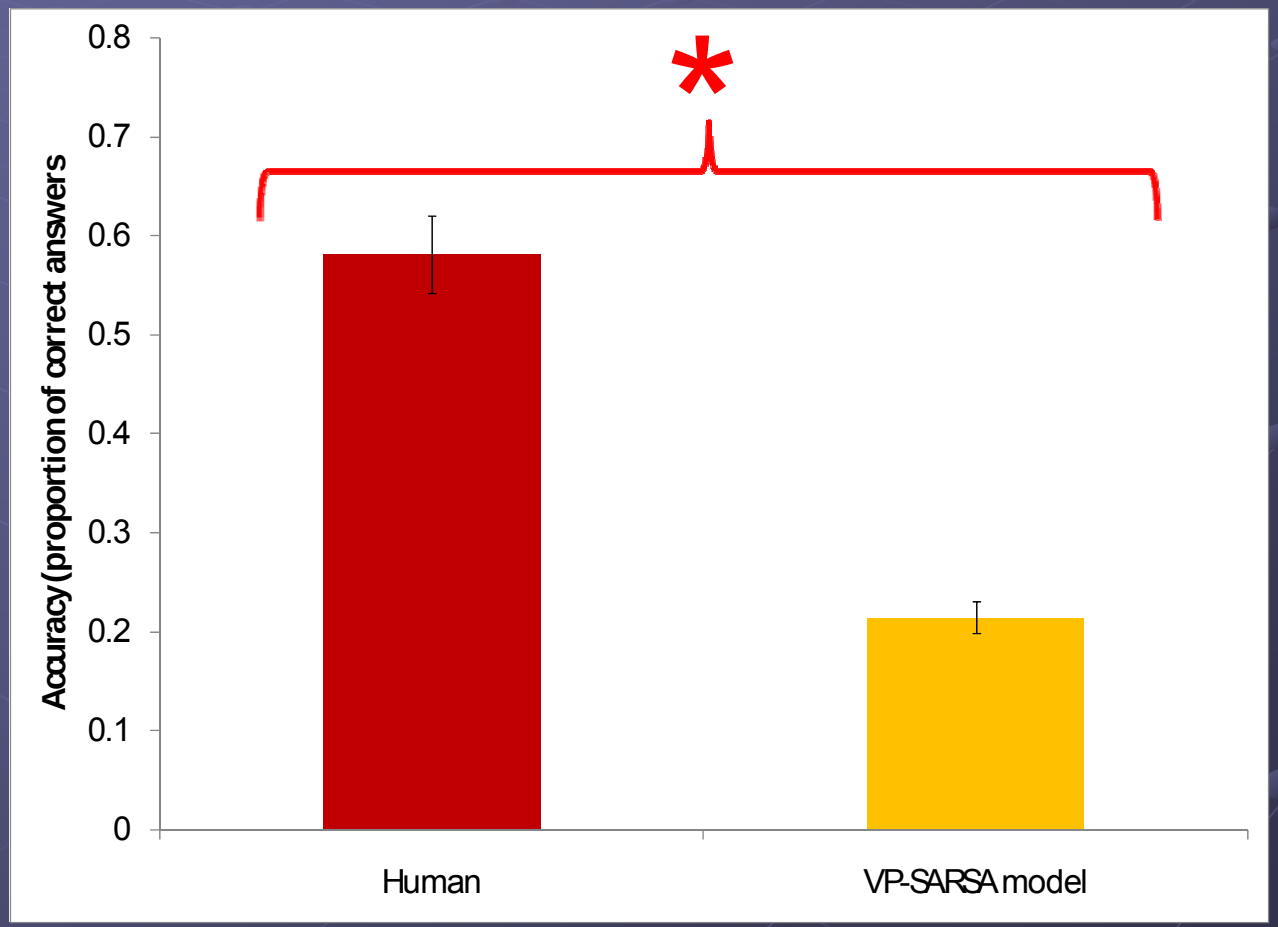
Results – Working memory

Working memory size	Networks trained to success (out of 20)	Successful networks (attain 100% accuracy)		Unsuccessful networks (no solution found in within 10000 epochs or 100 recruits)		
		Mean recruits	Mean epochs	Mean recruits	Mean reward	Mean accuracy
1	**0**	N/A	N/A	12.5	-2.9	0.44
3	14	16.4	1473	18.5	3.0	0.56
5	**19**	19.2	1537	31.0	-2.0	0.46
10	7	25.6	**4350**	**71.4**	-5.8	0.38

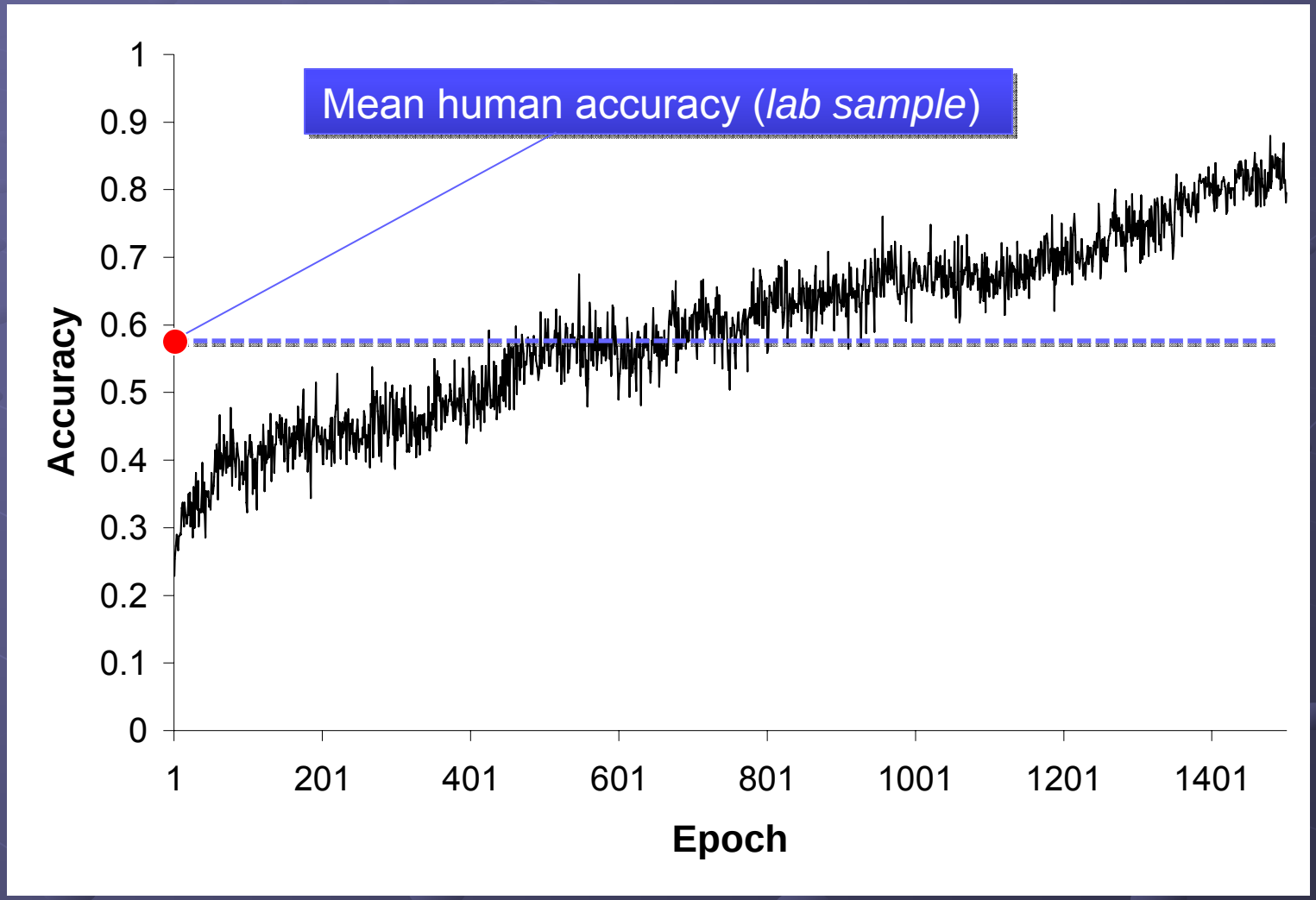
- Too few actions: stuck in local reward maximum
- Too many actions: gets lost searching for solution
- Optimal: compatible with working memory size estimates

Results: Accuracy

- Models less accurate than humans (lab sample)



Average accuracy (20 networks)



Reinforcement model summary

Criterion	Value prediction
Match human accuracy	

Reinforcement model summary

Criterion	Value prediction
Match human accuracy	
Can learn task with impoverished rewards given enough training	+

Reinforcement model summary

Criterion	Value prediction
Match human accuracy	
Can learn task with impoverished rewards given enough training	+
Optimal range of exploration compatible with human working memory estimates	+

Computational models summary

	Action prediction	Value prediction
Imitation learning	+	+
Match human accuracy	+	+
More accurate		+
Trains faster	+	
More compact models (fewer recruits)	+	
Reinforcement learning		+ (SARSA)
Match human accuracy	N/A	

- Learn by imitation and rewards
  - Possible in Value prediction models
  - Switch between DPT and SARSA

GENERAL DISCUSSION

Why are humans more accurate than reinforcement models?

- Reasoning and mental rehearsing
  - Humans can mentally play alternative actions
  - Models need to explore more
- Richer reward structure
  - Humans probably monitor distance to goal – use closeness to goal as a reward
  - Evidence found in Think aloud protocols pilot study

Integration of cognitive mechanisms

	Action prediction	Value prediction
Working memory		**+ (Softmax)**
Monitoring distance to goal		**Possible (Distance-Based Rewards)**
Search		**Possible**
Motivation		**Possible**

- Search – TD-leaf (Baxter, Tridgell, & Weaver, 1998)
- Motivation – intrinsically motivated reinforcement learning (Singh, Barto, & Chentanez, 2004)

Model comparison

	Action prediction	Value prediction
Cover human imitation accuracy	+	+
Problem representation allows alternative actions		+
Learning by imitation and reinforcement		+
Integration of multiple cognitive mechanisms in a single, parsimonious model		+
Requires less training data	+	
Model simplicity - Fewer modules	+	
Compact models that train faster	+	

Take home messages

- Experimental
  - Imitation & verbal instructions groups more accurate than reinforcement
  - Demos & instructions reduced simplicity bias
  - Use reasoning & monitoring distance to goal
  - Understand, not memorize, demonstrations
- Computational models
  - Cognitively plausible (Cascade-Correlation, SARSA)
  - Good coverage of human performance (except RL)
  - Value prediction models: Integration of cognitive mechanisms
  - Interesting alternatives to symbolic models

The End

- Acknowledgments - Collaborators
  - Thomas R. Shultz
  - Kristine H. Onishi
  - François Rivest
  - Simcha Samuel
  - Melissa Bowen
- Funding
  - NSERC (Natural Sciences and Engineering Research Council of Canada)
  - Lloyd Carr-Harris McGill major scholarship

- Comments, questions?

References (1)

- Adams, J. L. (1974). *Conceptual blockbusting: A guide to better ideas. Reading, MA: Addison-Wesley.*
- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, D., Lebiere, C., & Qin, Y. (2004). An integrated theory of mind. *Psychological Review, 111, 1036-1060.*
- Baxter, J., Tridgell, A., & Weaver, L. (1998). TDLeaf(lambda): combining temporal difference learning with game-tree search. In *the Proceedings of the ninth Australian Conference on Neural Networks (pp. 168-172).*
- Byrne, R. W., & Russon, A. E. (1998). Learning by imitation: A hierarchical approach. *Behavioral and Brain Sciences, 21(05), 667-684.*
- Carpenter, M., Call, J., & Tomasello, M. (2002). Understanding “prior intentions” enables two-year-olds to imitatively learn a complex task. *Child Development, 75(3), 1431-1441.*
- Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24, 87-125.*
- Dandurand, F., Shultz, T. R., & Rivest, F. (2007). Complex problem solving with reinforcement learning. In *the Proceeding of the 6th IEEE International Conference on Development and Learning (ICDL-2007) (pp. 157-162): IEEE.*
- Dandurand, F., Shultz, T. R., & Onishi, K. H. (2007). Strategies, heuristics and biases in complex problem solving. In *the Proceedings of the twenty-ninth meeting of the Cognitive Science Society (CogSci 2007) (pp. 917-922). New-York: Lawrence Erlbaum Associates.*

References (2)

- Dandurand, F., Samuel, S., & Shultz, T. R. (2007). Imitation learning in problem solving tasks: Memorizing or understanding? In *the Proceedings of Cognitio 2007. Newcastle upon Tyne, United Kingdom: Cambridge Scholars Publishing.*
- Dandurand, F., Bowen, M., & Shultz, T. R. (2004). Learning by imitation, reinforcement and verbal rules in problem solving tasks. In *the Proceedings of the third International Conference on Development and Learning (ICDL'04). La Jolla, California, USA: The Salk Institute for Biological Studies.*
- Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture. In *Advances in neural information processing systems 2, D. S. Touretzky (ed.) (pp. 524-532). Los Altos, CA: Morgan Kaufmann*
- Freyd, J., & Tversky, B. (1984). Force of symmetry in form perception. *American Journal of Psychology, 97* (1), 109-126.
- Glass, A. L., & Holyoak, K. J. (1986). *Cognition, 2nd edition*. New York, NY: Random House.
- Holyoak, K. J. (1995). Problem solving. In E. E. Smith & D. N. Osherson (Eds.), *Thinking: An Invitation to Cognitive Science, 2nd edition (Vol. 3, pp. 267-296). Cambridge, MA: MIT Press.*
- Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J.C.Houk, J. L. Davis & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia (pp. 249-270). Cambridge, MA: MIT Press.*

References (3)

- Katona, G. (1940). *Organizing and memorizing. New York, NY: Columbia University Press.*
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review 63, 81–97.*
- Newell, A., & Simon, H. (1963). GPS: A program that simulates human thought. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought. New York, NY: McGraw-Hill.*
- Newell, A., & Simon, H. A. (1972). *Human problem solving. Englewood Cliffs, NJ: Prentice-Hall.*
- Newell, A. (1990). *Unified theories of cognition. Cambridge, MA: Harvard University Press.*
- Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science, 310, 1337-1340.*
- Shultz, T. R. (2003). *Computational developmental psychology. Cambridge, MA: MIT Press.*
- Singh, S., Barto, A. G., & Chentanez, N. (2004). Intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing (Vol. 18).*
- Suri, R. E., & Schultz, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience 91, 871-890.*
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction. Cambridge, MA: MIT Press.*