Complex problem solving with reinforcement learning

Frédéric Dandurand

Department of Psychology, McGill University

Presented at ICDL 2007

London, UK, July 11-13, 2007

Problem solving

- Many activities of humans and robots involve planning and problem solving
- Classical research area in cognitive psychology
- Information processing theory (Newell & Simon, 1972) [1]
  - Emphasized search and heuristics (means-ends analysis)

Cognitive models of problem solving

- Current computational models
  - Largely symbolic (e.g., SOAR, ACT-R)
- New model: reinforcement-learning based

(Sutton & Barto, 1998) [2]

Research questions

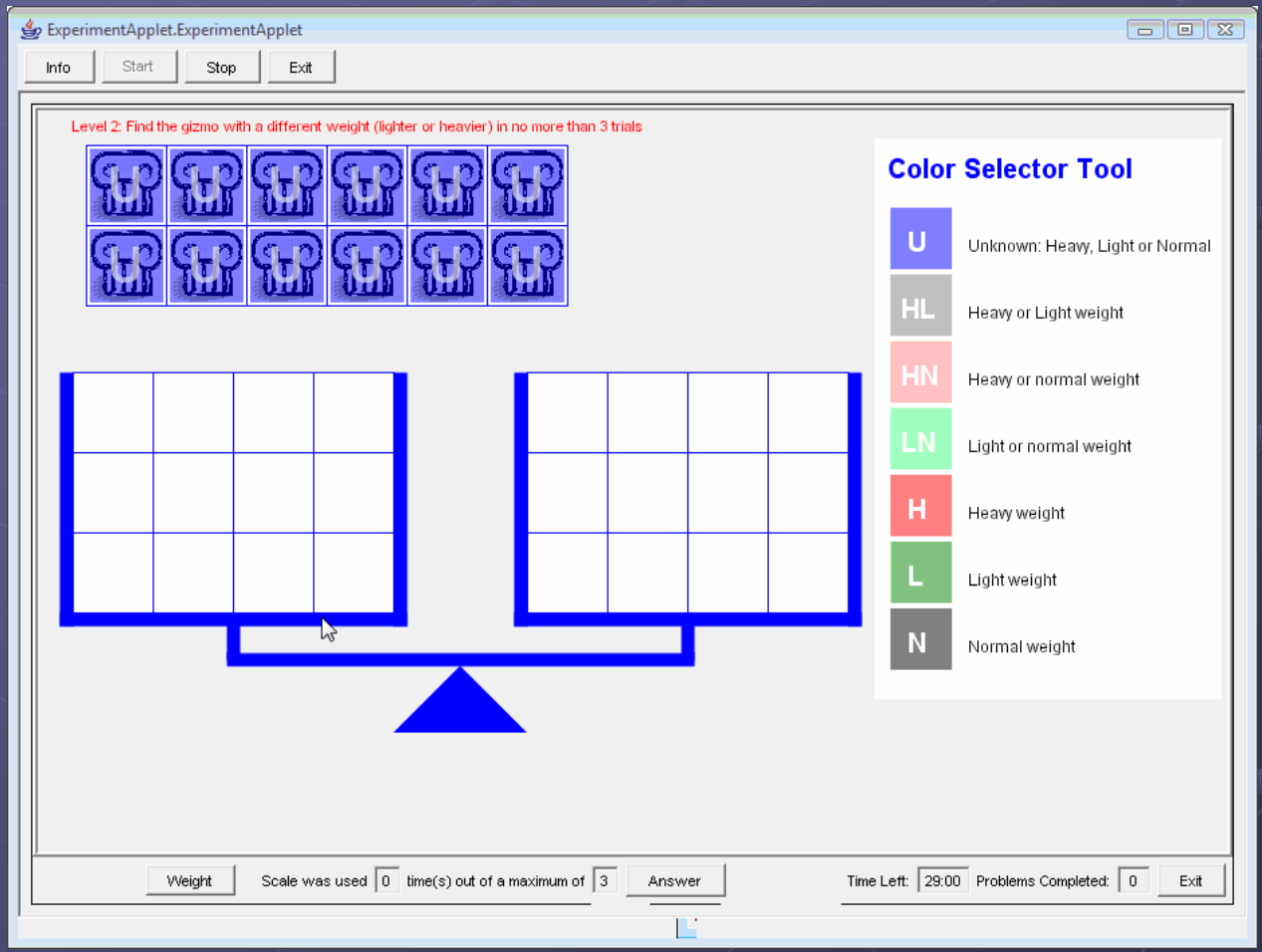
- Can reinforcement-based systems learn complex problem solving tasks?
- Does limiting exploration improve performance?
- Comparing humans and models
  - Can models capture human accuracy?
  - Can they model human biases?

Why reinforcement learning models?

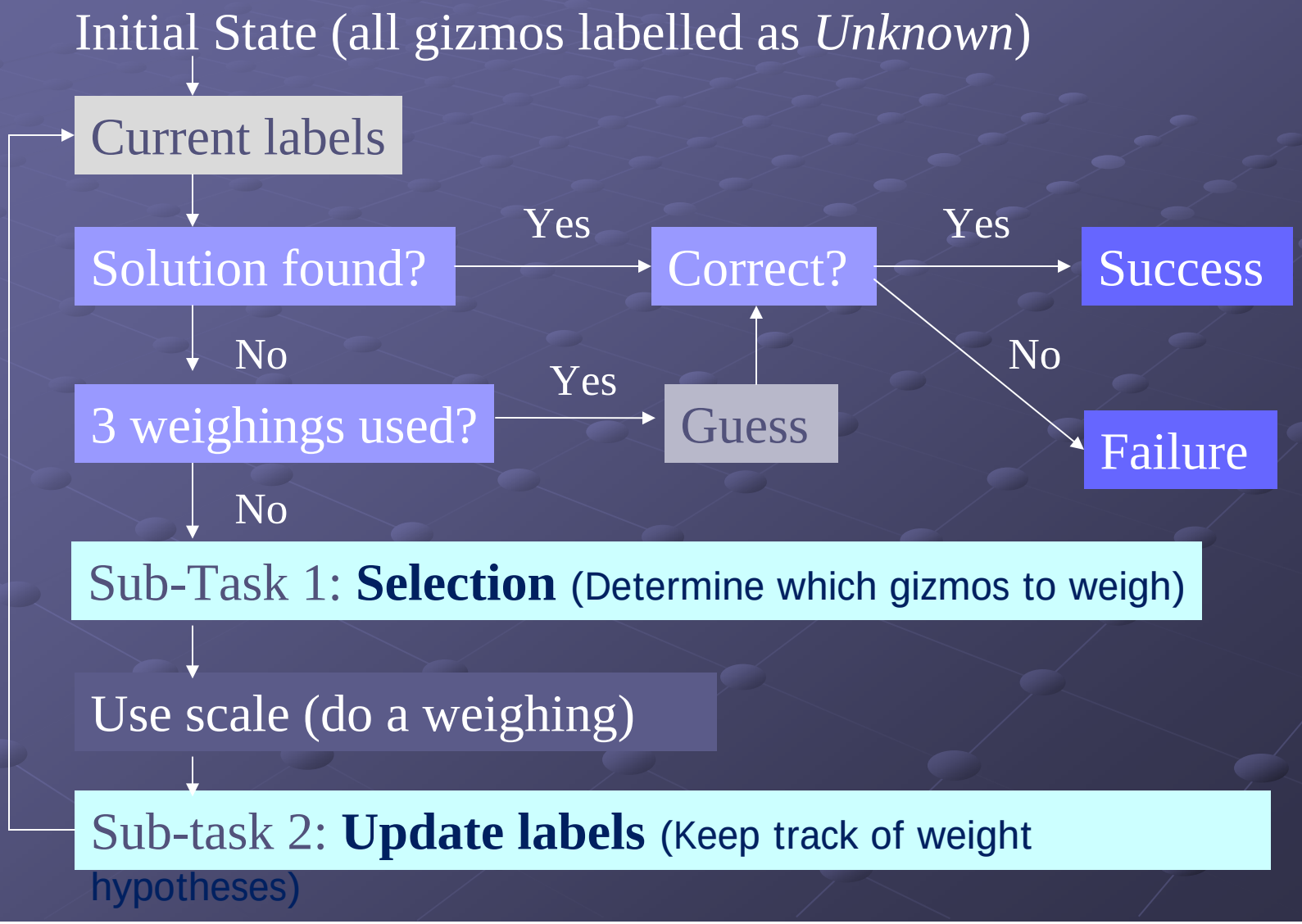
- Compatibility between reinforcement learning and information processing theory
  - States, transitions, constraints
- Biologically plausible
  - SARSA-like mechanisms using the basal ganglia, dopamine and the striatum [3, 4, 5]
- Previous successes
  - Psychology: classical & operant conditioning
  - Machine learning: Backgammon [6]

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



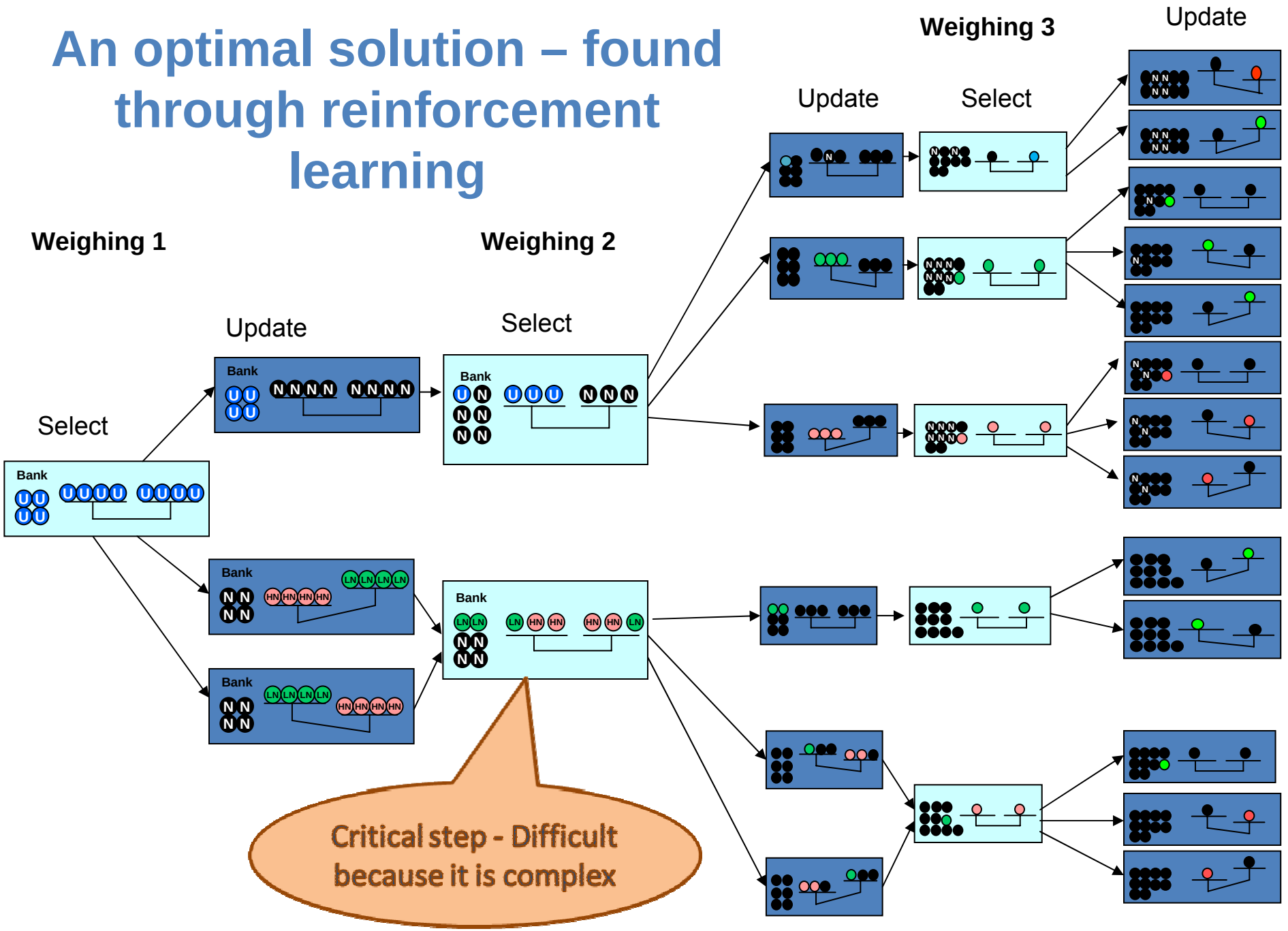
Task Analysis for humans



Problem characteristics

- 24 cases to discriminate (12 gizmos x 2 weights)
- Success criterion
  - Problem solved when strategy yields 100% accuracy – reliable solution of all cases
  - Sub-optimal solutions involve guessing

An optimal solution – found through reinforcement learning



Search space

- 6 187 states
- 5 671 402 selection actions
- Average of 916 selection actions per state

Human performance

- Completed an average of 17.0 trials in 30 minutes (Dandurand, Bowen & Shultz, 2004) [7]
- Accuracy (proportion of correct answers) = 59.0% (Dandurand, Bowen & Shultz, 2004) [7]
- No one found an optimal solution within 30 minutes (never found the critical step)

Reinforcement learning model

- Algorithm for learning expected rewards
- Reward structure
- Method to select an action
- System to generate expected rewards

Model simplifications

- Learn selection subtask only
  - If learning two tasks (selection and update) at the same time, solution is a moving target
  - Assume an ideal agent for update task
- Only consider weighings with equal numbers of gizmos on the two sides of the scale
  - Humans: equal weighings 98.6% of trials

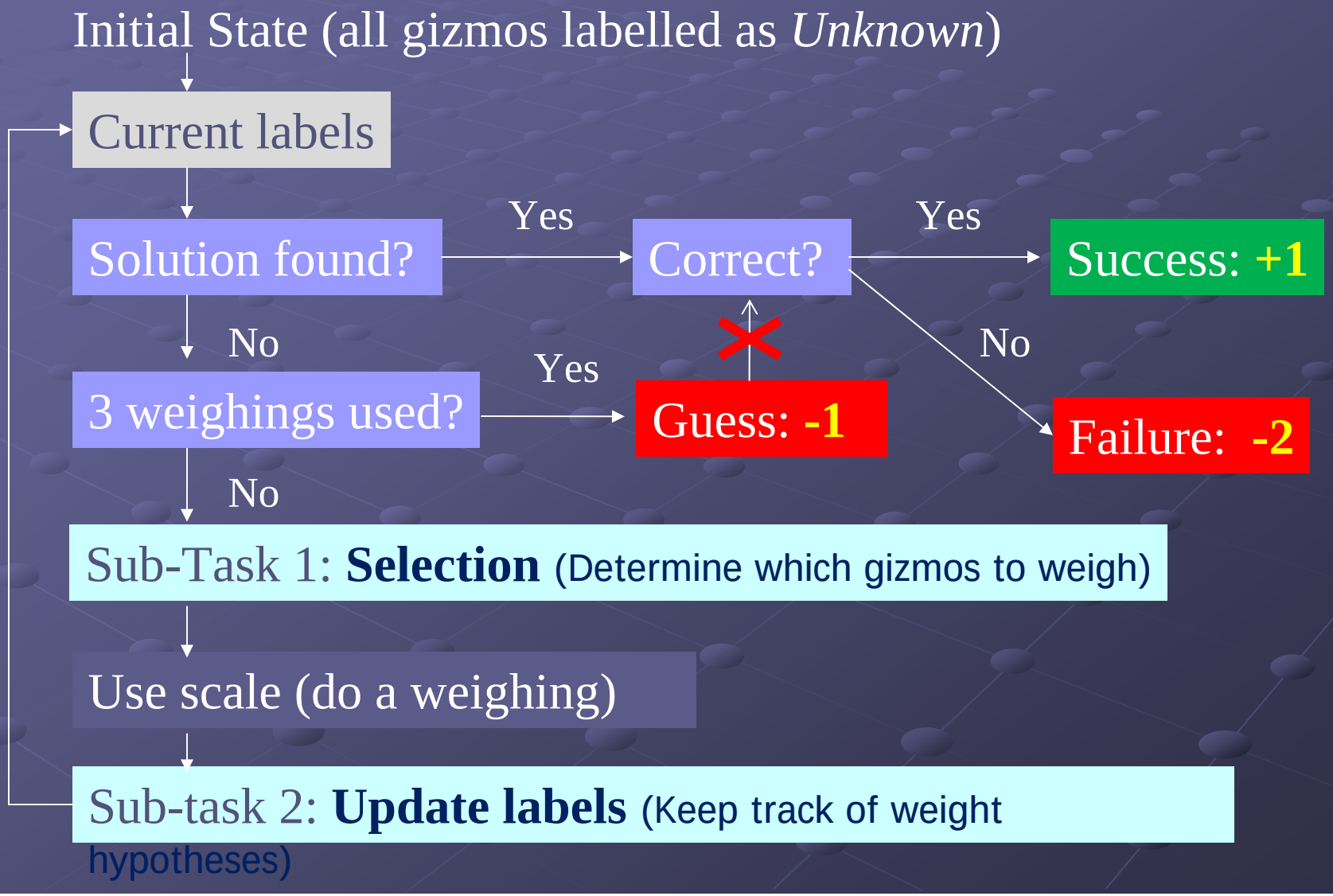
Learning expected rewards

- SARSA $(s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1})$ (Sutton & Barto, 1998) [2]

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

  - Q is the predicted reward
  - s is a state, a is an action, r is a reward
  - indices t and t+1 are used for current and next states and actions respectively
  - α is the learning rate (0.5)
  - γ is the discount factor (1.0)

Task analysis for model



Action selection method

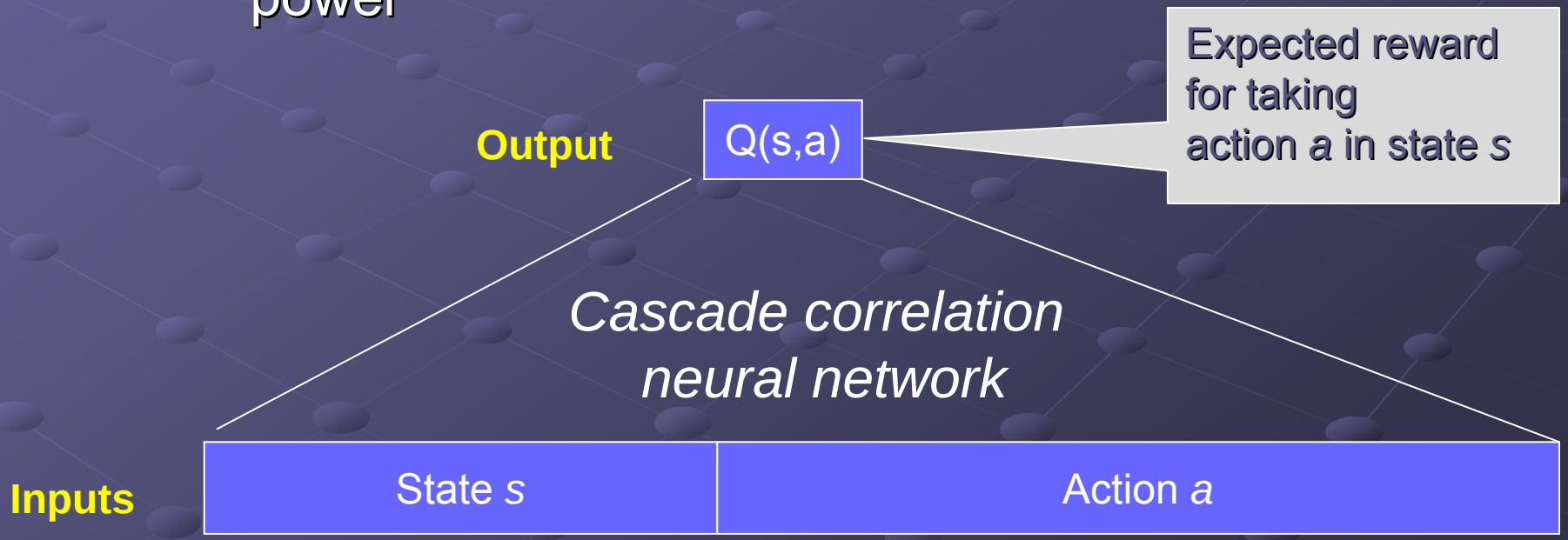
- Used a modified Softmax method
  - Softmax: higher expected value → higher probability of taking action
  - Keep only the *n* best actions (with highest predicted rewards) from a given state
- ≈ Working memory
  - Number of active elements

Computation of expected rewards

- Look-up table
  - Search space too large
- Why use function approximator?
  - Ability to generalize
  - Cognitively plausible

Function approximator

- Used cascade correlation neural networks
  (Fahlman, 1990) [8]
  - Constructive neural networks
    - Recruit hidden units to increase computational power



Experiment 1: 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) [9]
  - 4 +/- 1 (Cowan, 2000) [10]

Results – Working memory

Working memory size	Networks trained to success (out of 20)	Successful networks (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	0.56
5	**19**	19.2	1537	31	-2	0.46
10	7	25.6	**4350**	**71.4**	-5.8	0.38

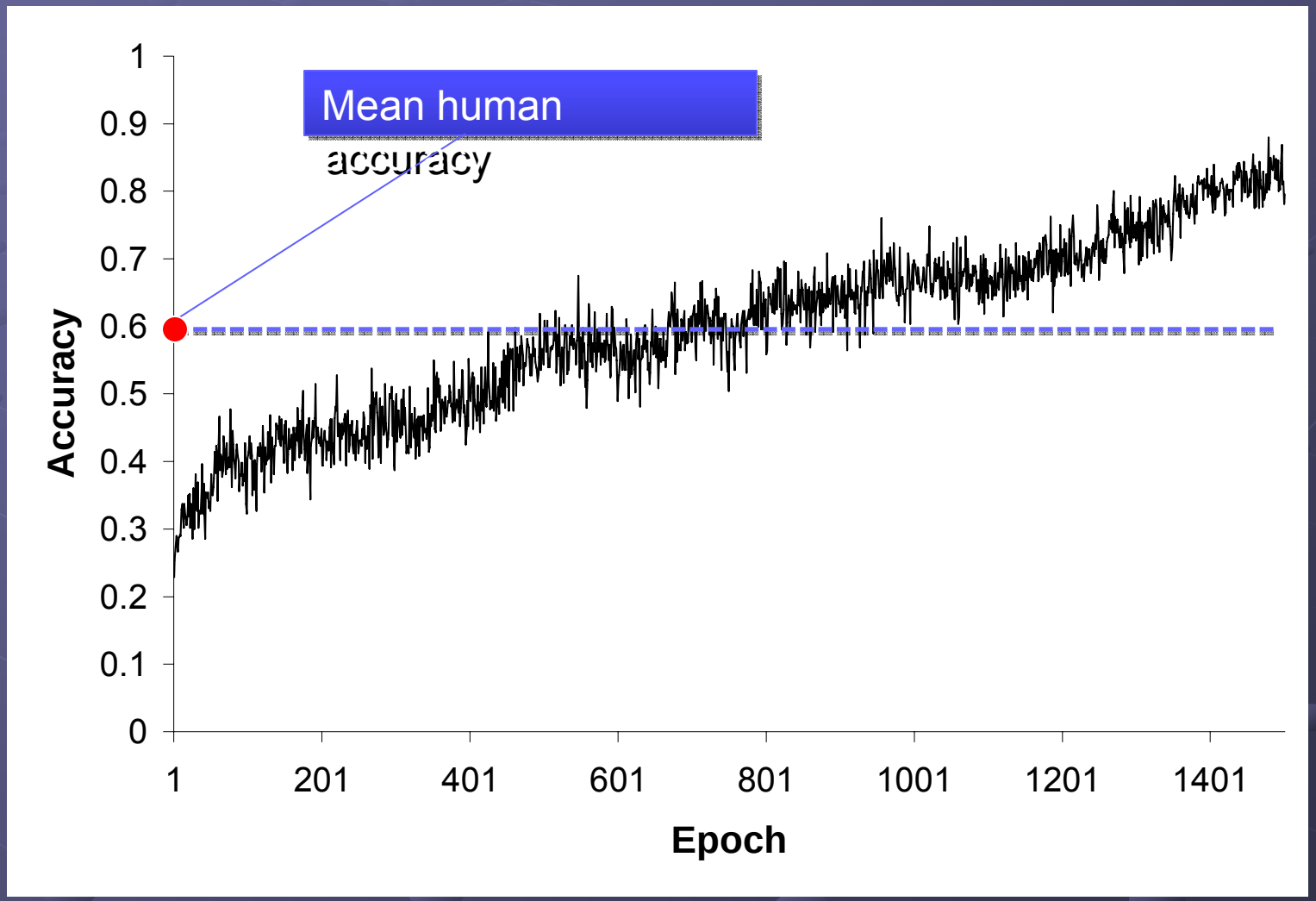
- Too few actions: stuck in local reward maximum
- Too many actions: gets lost searching for solution
- Optimal: compatible with working memory size estimates

Experiment 2: Human vs. model

	Training	Accuracy *M* *(sd)*	Selection complexity *M* *(sd)*
Human	17.0 trial	**0.59** (0.49)	**2.24** (0.31)
Model	1 epoch (24 episodes)	**0.21** (0.07)	**2.63** (0.37)

- Reinforcement model (working memory size = 5)
  - Less accurate than humans $F(1,39) = 70.8, p < 0.001$
  - More complex $F(1,38) = 17.2, p < 0.001$

Average accuracy (20 networks)



Experiment 3: Simplicity bias

- Humans prefer simple solutions

  (Dandurand, Shultz & Onishi, 2007) [11]

- Pre-trained model to prefer simplicity
  - Penalized selection complexity

$$Q_{penalized} = Q_{init} - 0.25 \times \text{complexity index}$$

Selection complexity

	Accuracy *M* *(sd)*	Selection complexity *M* *(sd)*
Human	**0.59** (0.49)	**2.24** (0.31)
Model with no bias	**0.21** (0.07)	**2.63** (0.37)

Selection complexity

	Accuracy *M* *(sd)*	Selection complexity *M* *(sd)*
Human	**0.59** (0.49)	**2.24** (0.31)
Model with bias	**0.21** (0.03)	**2.29** (0.36)
Model with no bias	**0.21** (0.07)	**2.63** (0.37)

- Pre-trained networks captured human bias for simplicity
- No effect on model accuracy

Results – Summary

- Reinforcement-based models can learn the Gizmo problem solving task
- Intermediate range of exploration was optimal
- Comparing humans and models
  - Models are less accurate than humans
  - Models capture human simplicity bias

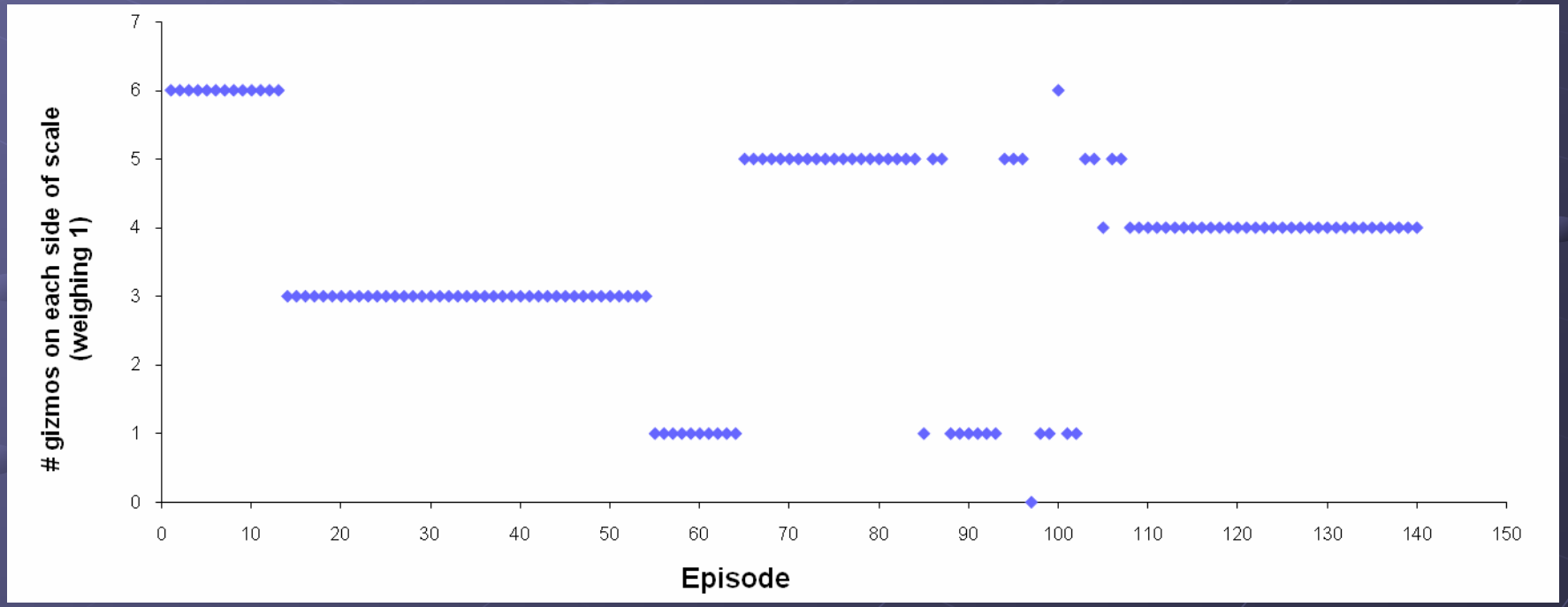
Why are humans more accurate than models?

- 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

Exploration

- Models may need to explore more than humans because models use no reasoning and have access to poorer reward structure

Exploration example (1st weighing)



Next steps

- What cognitive mechanisms do humans use?
  - Think aloud protocols
- Computational models
  - Reasoning: TD-leaf (Baxter, Tridgell & Weaver, 1998) [12]
  - Distance-Based Rewards (DBR)
    - Implement means-ends analysis
    - Use closeness to goal as reward

Take home messages

- Humans use more than only rewards
- Reinforcement-based system are promising for modeling human problem solving

Thank you !

- Acknowledgments
  - Collaborators
    - Thomas R. Shultz and François Rivest
  - Comments and suggestions
    - Kristine H. Onishi and Kayo Nakamura
  - Funding
    - NSERC (Natural Sciences and Engineering Research Council of Canada)
    - Lloyd Carr-Harris McGill major scholarship

- Comments, questions?

References

- [1] A. Newell and H. A. Simon, *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall, 1972.
- [2] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, A Bradford Book, 1998.
- [3] K. Samejima, Y. Ueda, K. Doya, and M. Kimura, "Representation of action-specific reward values in the striatum," *Science,* vol. 310, pp. 1337-1340, 2005.
- [4] R. E. Suri and W. Schultz, "A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task," *Neuroscience* vol. 91, pp. 871-890, 1999.
- [5] J. C. Houk, J. L. Adams, and A. G. Barto, "A Model of How the Basal Ganglia Generate and Use Neural Signals that Predict Reinforcement," in *Models of Information Processing in the Basal Ganglia* J.C.Houk, J. L. Davis, and D. G. Beiser, Eds.: MIT Press, 1995, pp. 249-270.
- [6] G. J. Tesauro, "Temporal difference learning and TD-Gammon," *Communications of the ACM* vol. 38, pp. 58-68, 1995.
- [7] F. Dandurand, M. Bowen, and T. R. Shultz, "Learning by Imitation, Reinforcement and Verbal Rules in Problem Solving Tasks," in *Third International Conference on Development and Learning (ICDL'04) Developing Social Brains*, La Jolla, California, USA, 2004, p. 26.
- [8] S. E. Fahlman and C. Lebiere, "The cascade-correlation learning architecture," *Advances in neural information processing systems 2, D. S. Touretzky (ed.),* pp. 524-532, 1990.
- [9] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychological Review* vol. 63, pp. 81–97, 1956.
- [10] N. Cowan, "The magical number 4 in short-term memory: A reconsideration of mental storage capacity," *Behavioral and Brain Sciences,* vol. 24, pp. 87-125, 2000.
- [11] F. D. Dandurand, T. R. Shultz, and K. H. Onishi, "Strategies, Heuristics and Biases in Complex Problem Solving," in *CogSci*, 2007, in press.
- [12] J. Baxter, A. Tridgell, and L. Weaver, "TDLeaf(lambda): Combining Temporal Difference Learning with Game-Tree Search," in *In Proceedings of the Ninth Australian Conference on Neural Networks*, Brisbane QLD, 1998, pp. 168-172.