Building Environments for Simulation & Experimentation in Malmo

David M. Schwartz and Dr. Christopher L. Dancy II

7/19/19 – ACT-R Workshop 2019

Not this Malmo…

Microsoft's Project Malmo

Introduction

- Cognitive architectures are models of the mind
  - Used to make simulations of human behavior
- For instance, to study behavior in dangerous environments
- Don’t simulate world (task environment) around a person

Simulations of Task Environments

- Domain specific
  - Interaction with a car is different from a computer
- Time consuming to develop
- May differ from reality
- Divert resources away from studying cognition

Purpose

- Goals:
  - Develop a tool to create and run experiments in various environments
- Aim to accomplish by:
  - Connecting the ACT-R cognitive architecture to Malmo

Part 1: Connecting ACT-R and Malmo

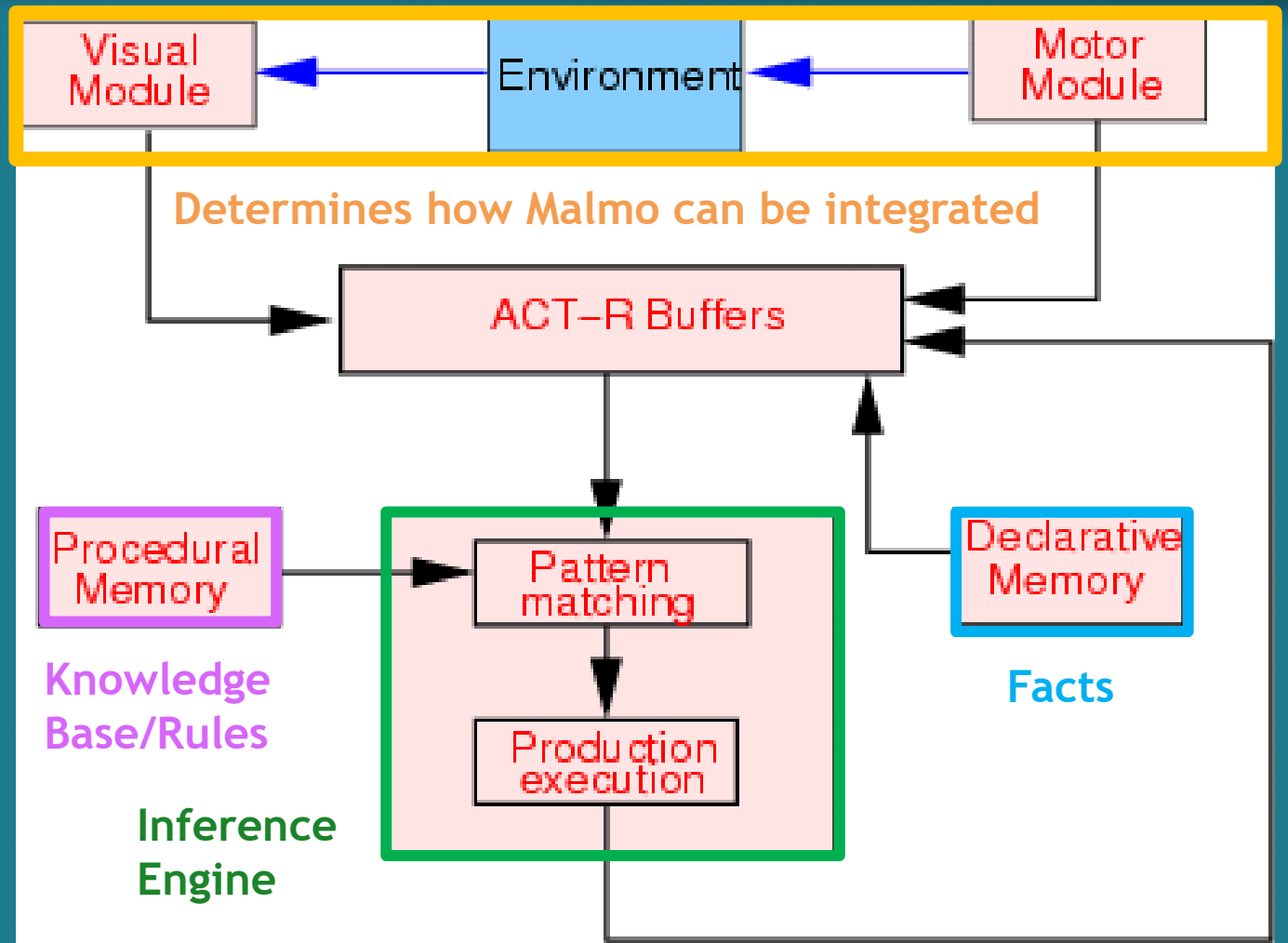
Project Malmo - Introduction

- Test suite for artificial general intelligence
- Uses Minecraft as an environment, providing:
  - Modifiable 3D world
  - Items
  - Creatures

Project Malmo - Usage

- Create a mission
  - Allows you to modify game world and properties
  - Specify game data to collect
- Agents connect through a network interface
  - Receive observations detailing the game state
  - Send commands causing their character to act

ACT-R – System Diagram

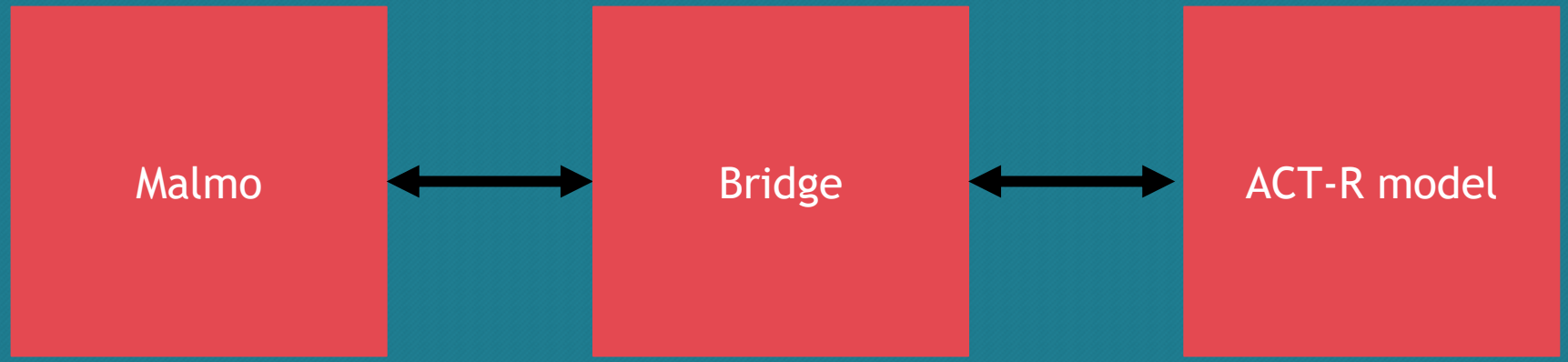


Compatibility

- Motor control in ACT-R and actions in Malmo go well together
  - Malmo commands based on game input
- Vision system also correlates well
  - World contains semantic blocks easily represented by symbols
- However, translation scheme necessary
  - Due to differences in representation

Diagram – Initial Bridge

- Son Pham and Dr. Dancy started the project
- Bridge updates experiment, collects data, and translates messages
- Takes advantage of Malmo and JSON Network Interface (JNI) Python APIs



[Some of the] Issues

- Bridge built on outdated software
  - Works with Python 2 and ACT-R 6
- Asynchronous
  - Big issue for data collection
  - Non-real time mode in ACT-R doesn't easily map onto Malmo

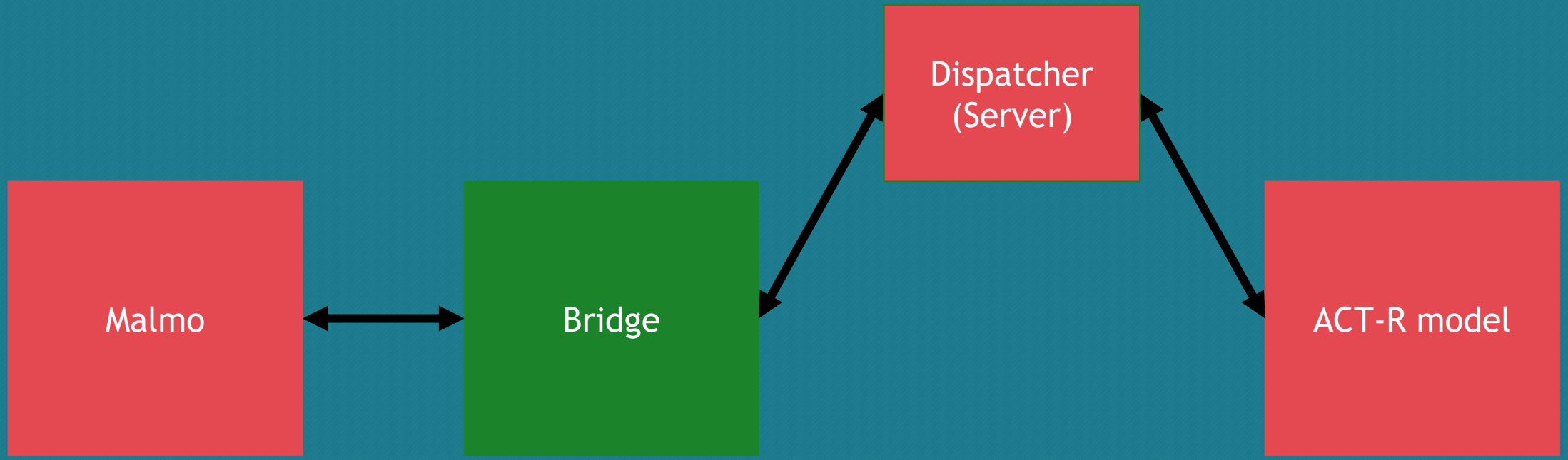
ACT-R 7.6+

- Modified to facilitate easier interaction with other programs
- JNI allowed other programs to send/receive events over a network
- 7.6+ takes idea to the next level
  - Remote procedure call based
  - All events are sent to a sever which executes them
  - Any program can send and monitor events

ACT-R 7.6+ - Implications

- Can access ACT-R clock from bridge (solves data collection issue)
- Synchronize components
  - ACT-R presses a button to move character forward
  - Bridge suspends execution of events until Malmo updates
  - Resumes normal operation

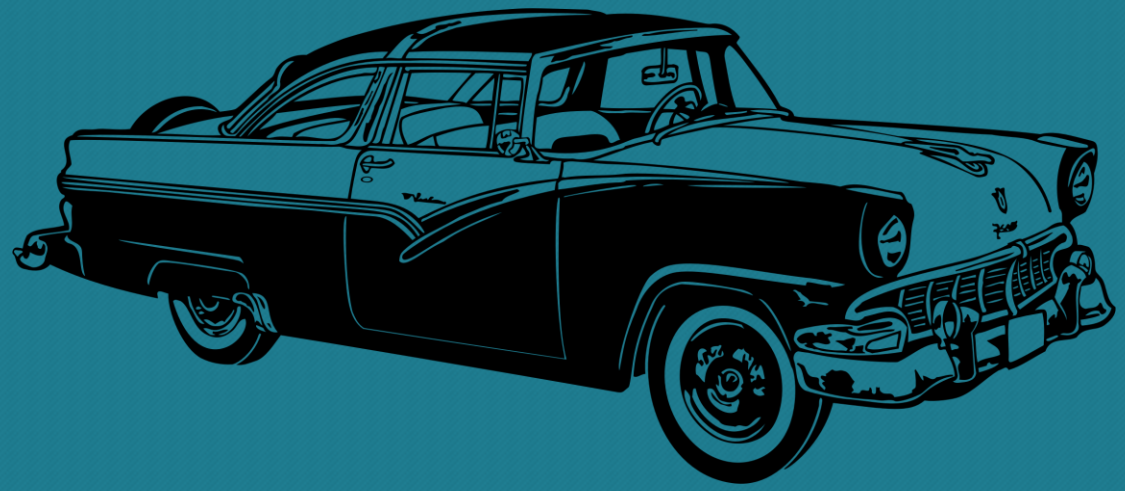
Diagram – New Bridge



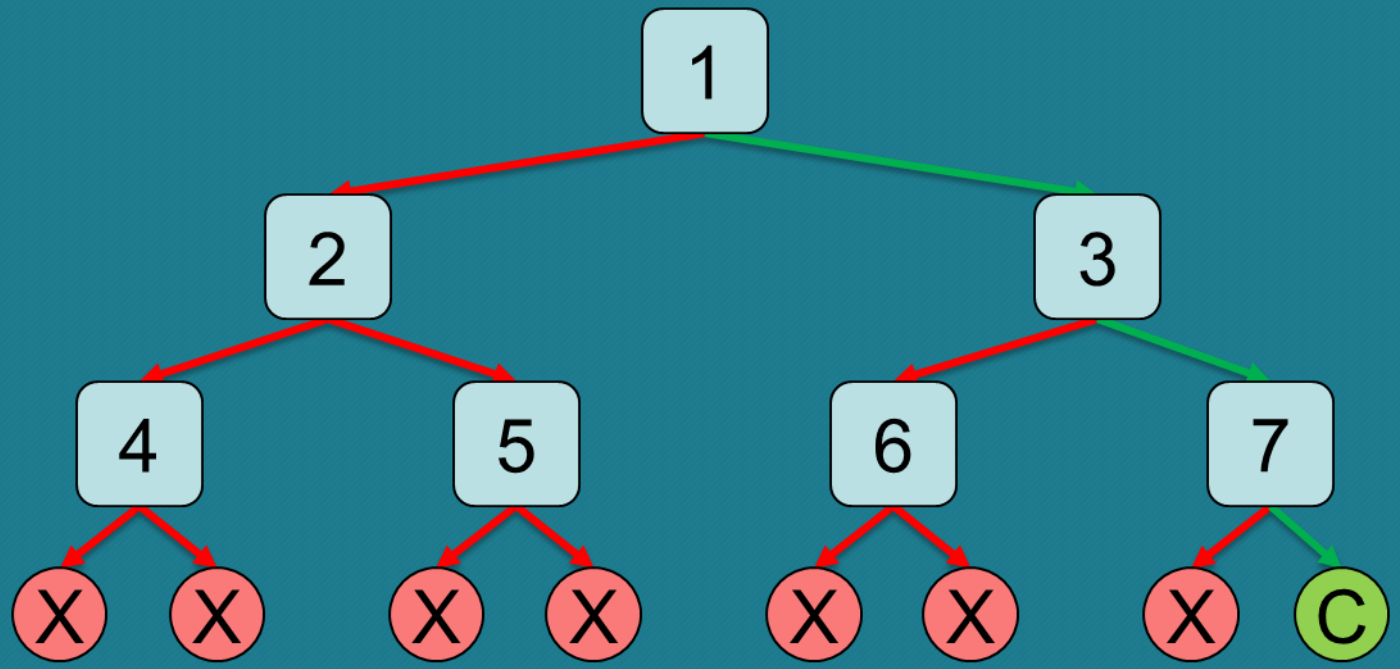
Part 2: Testing it
(whilst studying exploration and exploitation…)

Exploration Exploitation Trade-off

- Big role in our everyday lives
  - I.E. Buying a car, how many do you test drive?
- Cognitive model should manage it as well



Symbolic Maze



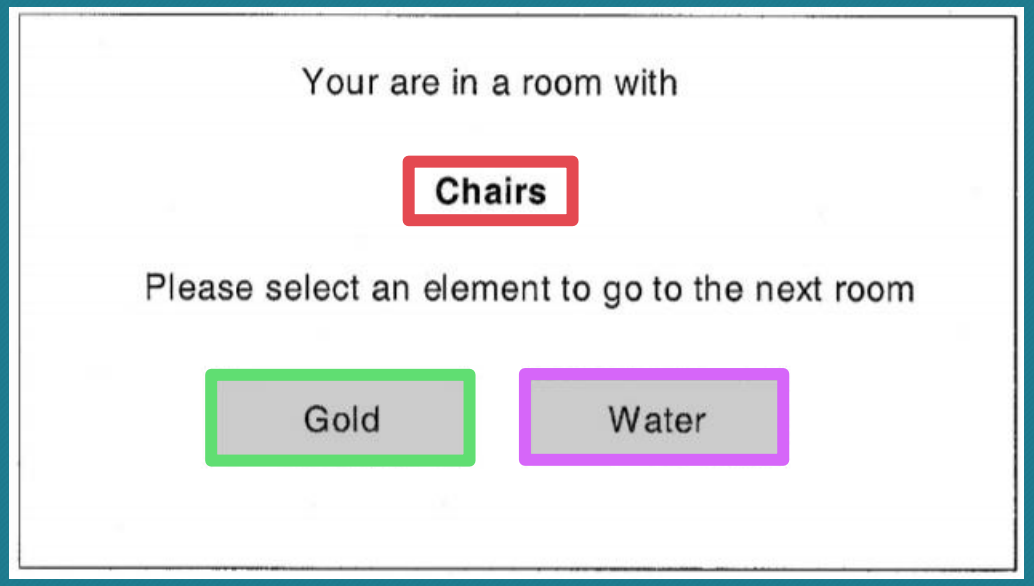
Your are in a room with

Chairs

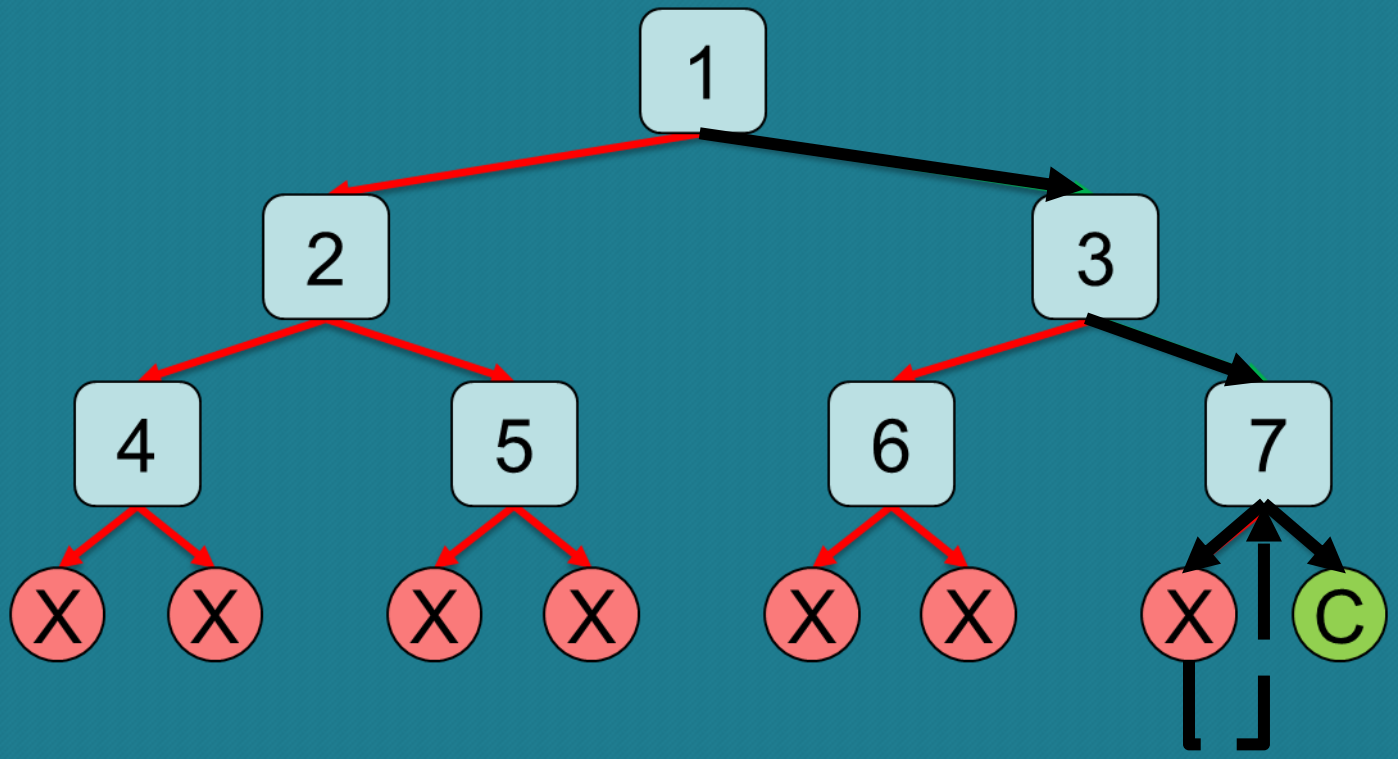
Please select an element to go to the next room

Gold | Water

Symbolic Maze – In Malmo

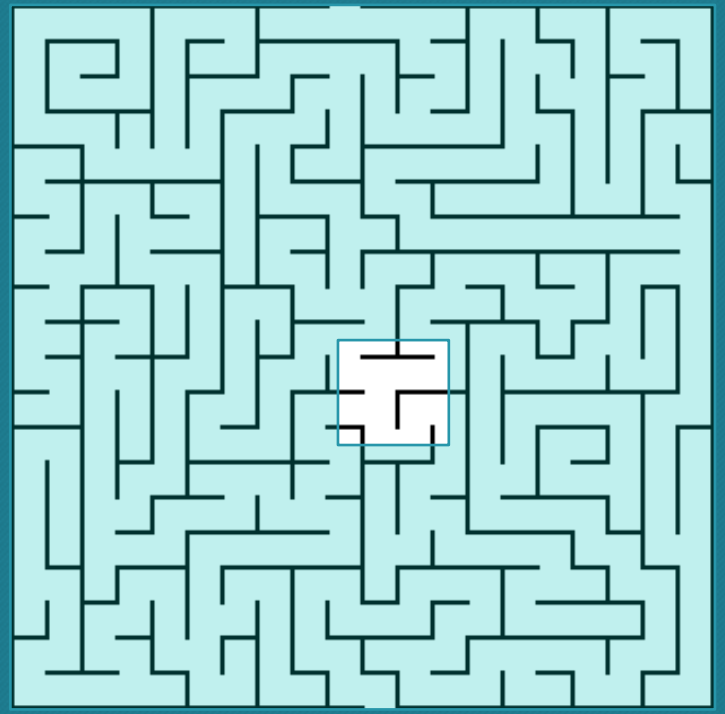


Symbolic Maze – Test Experiment



Symbolic Maze – Room Configurations

- Four different stimuli associated with one room
- Adds complexity to maze
- Equivalent to turning a paper maze



Symbolic Maze - Scoring

- In each round, player starts with three points
- Docked a point for each reset
  - Can’t go below zero
- Experiment score is the sum of scores for 200 rounds

Changing Environment

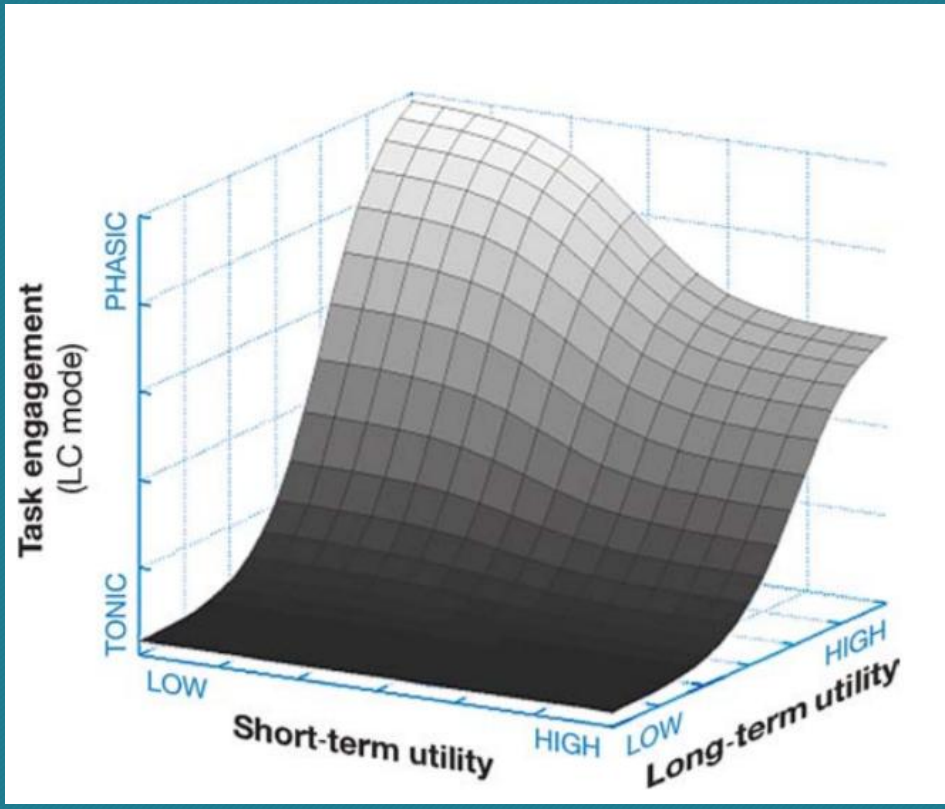
- Shows transition from exploring to exploiting
- Want to show inverse transition as well
  - Alter maze after 20 rounds
- Forces player to reconsider knowledge

How ACT-R Manages the Trade-off

- Utility learning – based on past experience
- Fails at reversal experiments
- Typical solution: annealing
  - Increase randomness is decision making
  - Model will eventually pick out a better behavior
  - Requires knowledge of when new approach needed

Adaptive Gain Theory

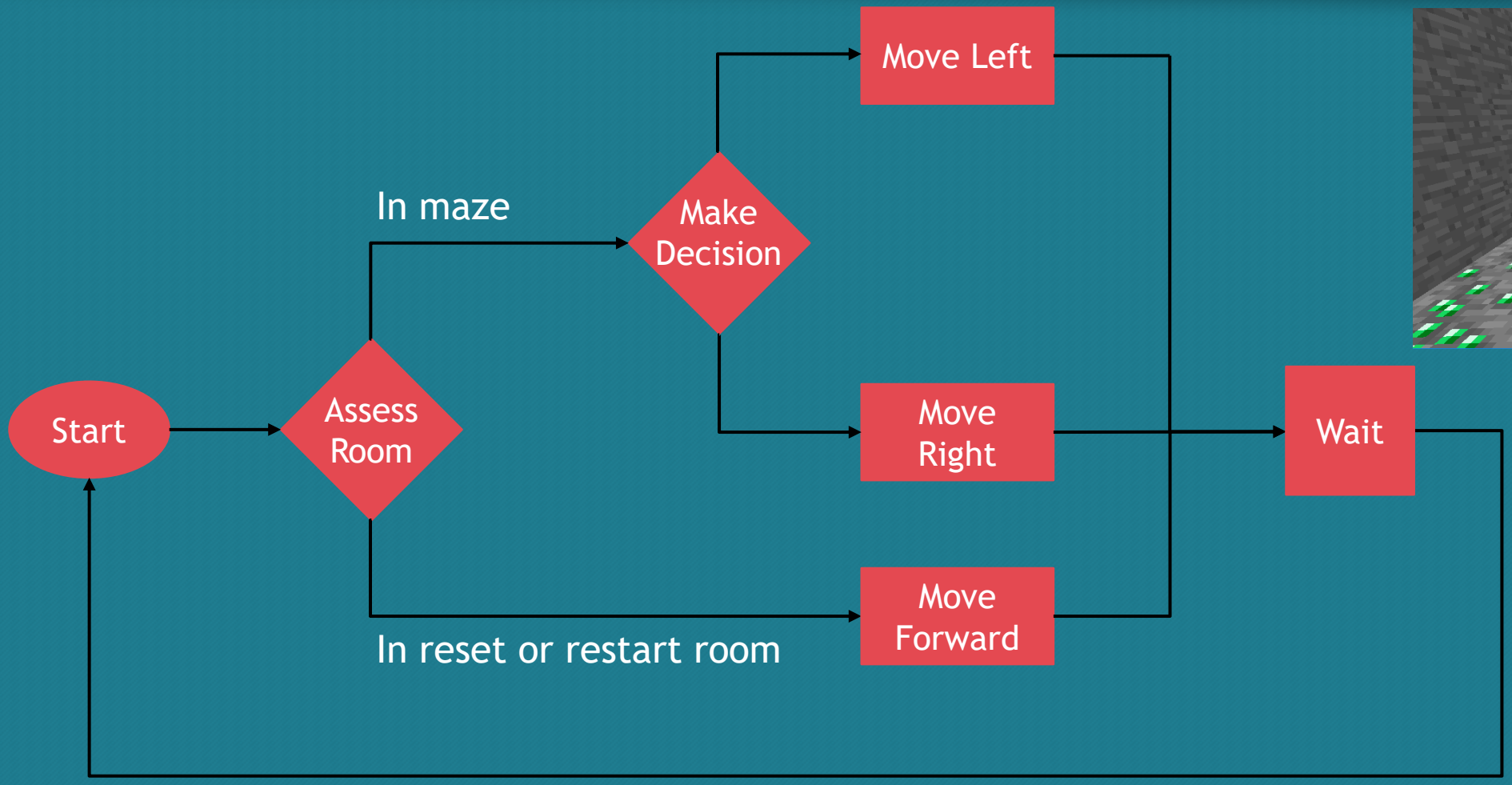
- About how Locus Coeruleus and Norepinephrine selectively filter information
- Related to task engagement
- Engagement (E) related to utility (U)
  - $E = \left(1 - logistic(U_{short})\right) * \left(logistic\left(U_{long}\right)\right)$



Modulated Annealing

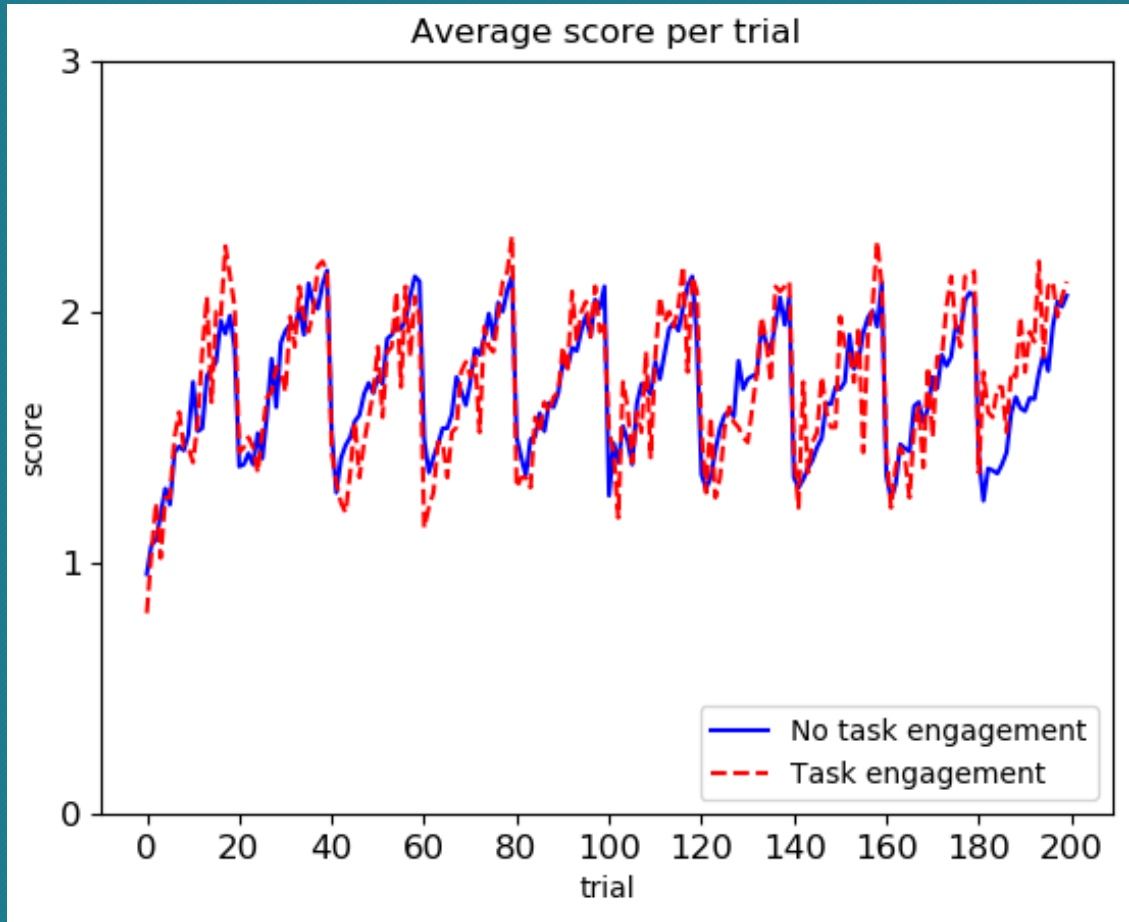
- Use task engagement to modulate temperature
  - Temperature refers to the amount of noise present
- Change temperature (T) according to engagement (E)
  - $T = (1 - E) * T_{max}$
- Automatically activates annealing and controls its magnitude

Model Flowchart



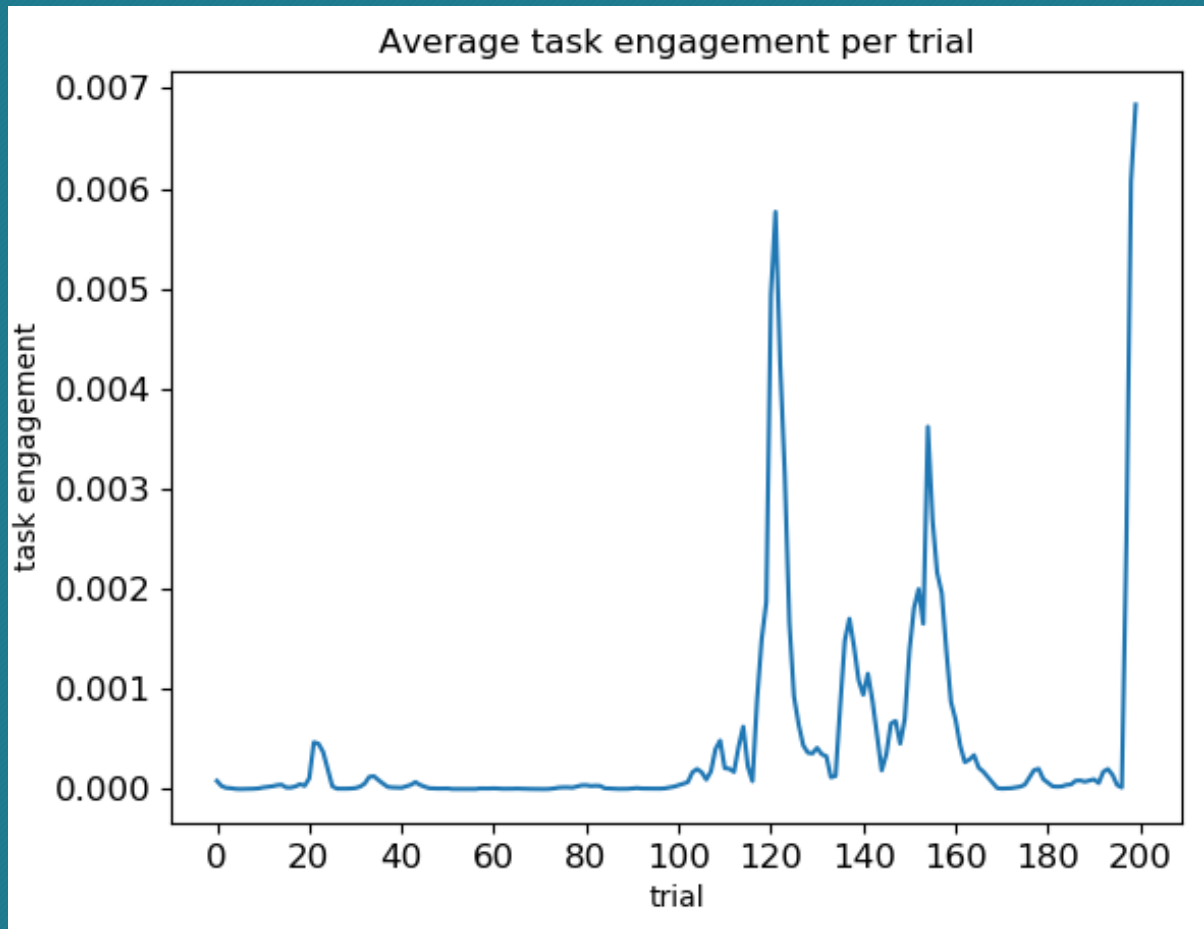
Results

- 250 simulations without task engagement
  - Parameters from Fu & Anderson (2006)
- 50 simulations with task engagement
  - Reduced number due to a bug
  - Short term utility over 3 seconds
  - Long term utility over 120 seconds
  - $T_{max} = 1$



Investigation

- Task engagement changed
  - Noise changed
- Noise settled to value of no task engagement model
- Caused by poor choice of parameters



Next Steps - Model

- Parameter sweep
  - Determine better starting values for task engagement
- Model learning
  - Current model uses a heuristic
- Collect subject data to compare to

Next Steps - Bridge

- Implement a standard control scheme
  - Researcher defines it in template
- Multiagent missions
- Utilize new Malmo API (which is under development)

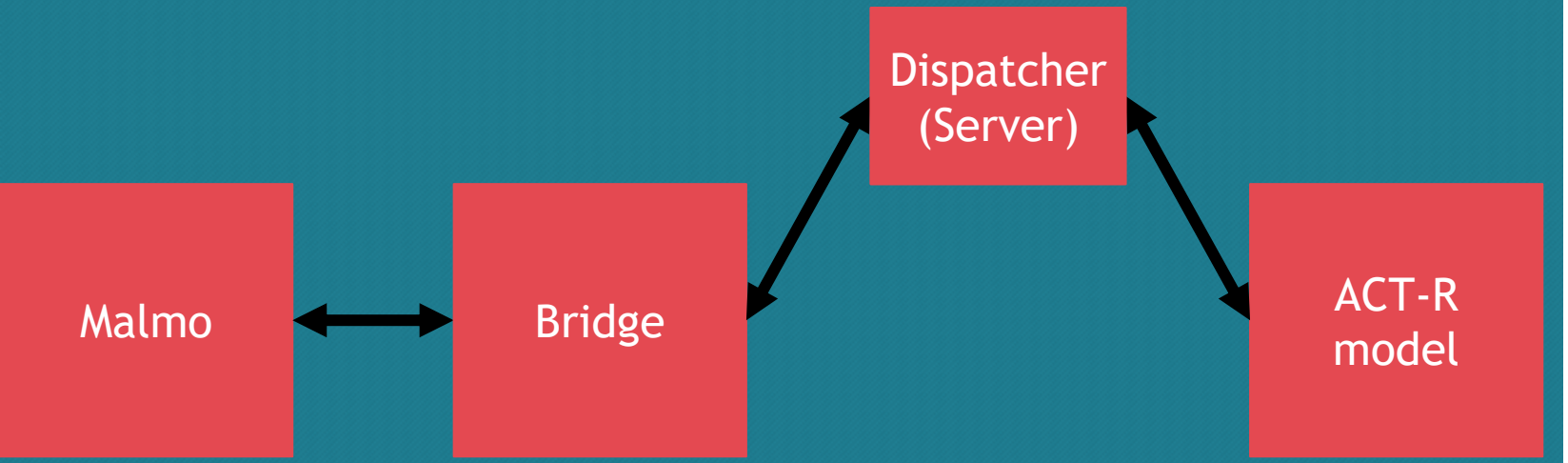
Conclusion

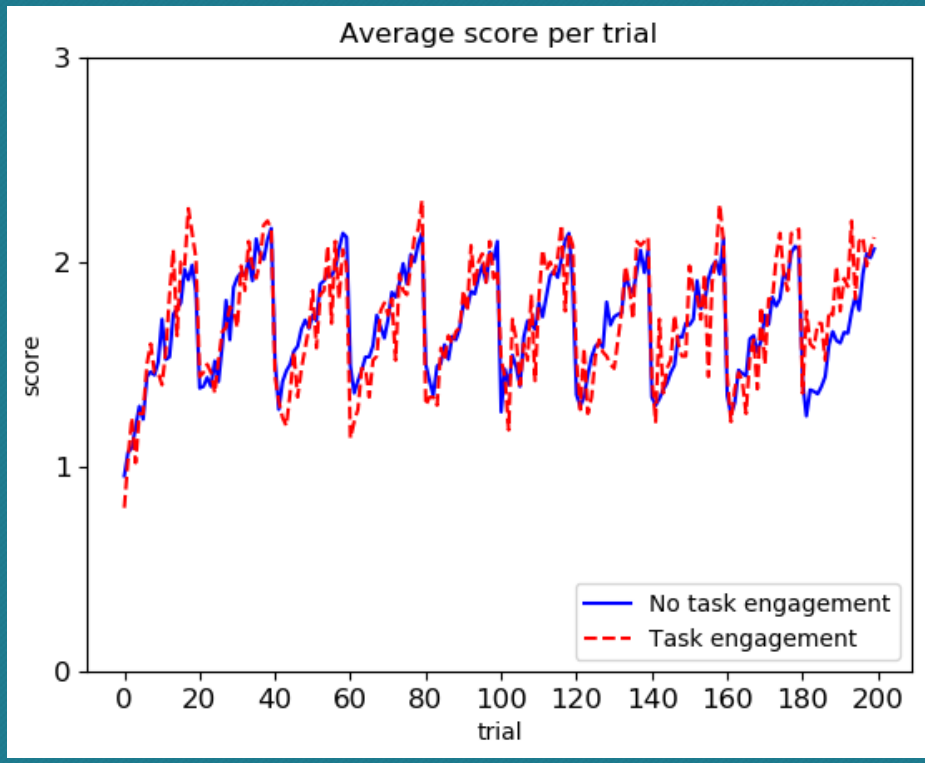
- Can create better simulations by incorporating the world around us
- Such simulations are time consuming to make
  - Bridge was made between ACT-R and Project Malmo
- Cognitive models must be able to explore and exploit its environment
  - Adaptive gain theory added to ACT-R

Resources

- Anderson, J.R. (2007). How Can the Human Mind Occur in the Physical Universe? *Oxford University Press*.
- Aston-Jones, G., & Cohen, J. (2005). An Integrative Theory of Locus Coeruleus-Norepinephrine Function: Adaptive Gain and Optimal Performance. *Annual Review of Neuroscience*.
- Cohen, J., McClure, S.M., and Yu, A.J. (2007) Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B.*
- Fu, W., & Anderson J. R. (2006). From recurrent choice to skill learning: A reinforcement-learning model. *Journal of Experimental Psychology: General,* 135(2), 184-206.
- Hope, R.M, Schoelles, M.J, and Gray, W.D. (2013). Simplifying the interaction between cognitive models and task environments with the JSON Network Interface. *Behavior Research Methods*. Vol 45.
- Johnson, M., Hofmann, K., Hutton, T., & Bignell, D. (2016). The Malmo platform for artificial intelligence experimentation. *In proceedings of Twenty-Fifth International joint conference on artificial intelligence (IJCAI),* New York, NY, 4246-4247.

Thanks for your time, questions?





Adaptive Gain Theory

- Theory about exploration exploitation trade-off
- Locus Coeruleus (LC) and Norepinephrine (NE) selectively filters information
- Two modes of distribution:
  - Phasic – filtering – exploitation
  - Tonic – no filtering - exploration

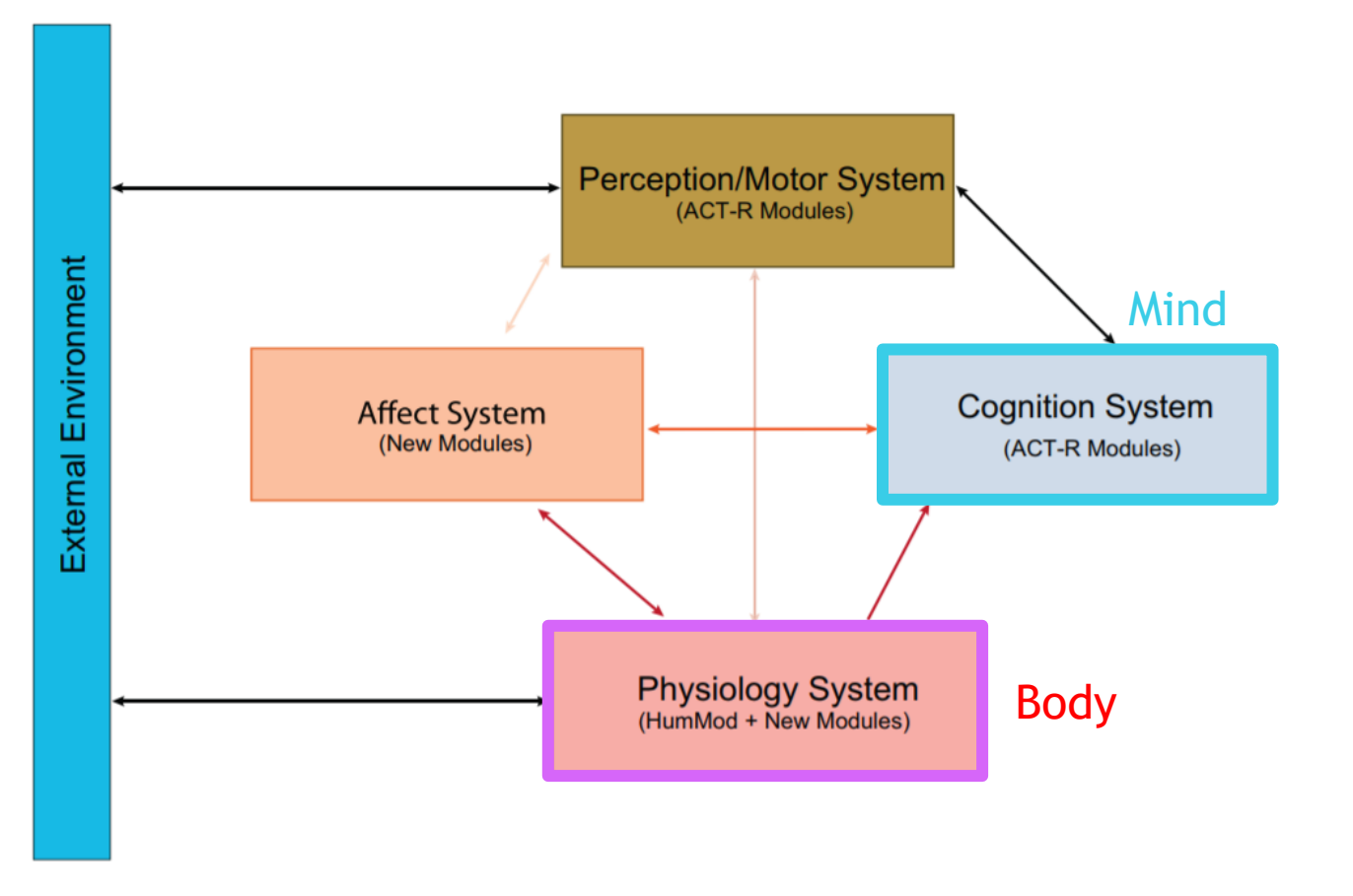
ACT-R Decision Making

- To pick a rule, ACT-R looks at:
  - Applicability – conditions that must be met
  - Utility – usefulness determined by experience
    - Based on reinforcement learning (Temporal Difference)
  - Noise – randomness, like a whim

Problem: Change

- Utility learning fails at reversal experiments
- Typical solution: annealing
  - Increase randomness is decision making
  - Model will eventually pick out a better behavior
  - Requires knowledge of when new approach needed

ACT-R/Φ


External Environment
Perception/Motor System
(ACT-R Modules)
Mind
Affect System
(New Modules)
Cognition System
(ACT-R Modules)
Physiology System
(HumMod + New Modules)
Body

Physiology System

- Hummod – Physiological simulation system
- Hormone generation and regulation
  - Cortisol, epinephrine, etc
- Why do we need it?
  - Disrupting bodily processes has a clear effect on decision making and recollection
    - IE you do better on a test when you have slept the night before

ACT-R Utility Learning

- Utility is based on reinforcement learning
- Temporal Difference Equation:
  - $U_i(n) = U_i(n-1) + \alpha[R_i(n) - U_i(n-1)]$
  - $U_i\ (n)$ is the utility of the *i*th production after its *nth* use
  - α learning rate parameter
  - $R_i\ (n)$ time diminished reward