2018

Towards a physio-cognitive model of the exploration exploitation trade-off.

David M. Schwartz
Bucknell University, dms061@bucknell.edu

Christopher L. Dancy
Bucknell University, cld028@bucknell.edu

Towards a Physio-Cognitive Model of the Exploration Exploitation Trade-off

David M. Schwartz (dms061@bucknell.edu), Christopher L. Dancy (christopher.dancy@bucknell.edu)
Department of Computer Science, Bucknell University
701 Moore Avenue,
Lewisburg, PA 17837 USA

Keywords: exploration vs exploitation; utility, ACT-R/Φ, Project Malmo, reinforcement learning.

Introduction

Managing the exploration vs exploitation trade-off is an important part of our everyday lives. It occurs in minor decisions such as choosing what music to listen to as well as major decisions, such as picking a research direction to pursue. The dilemma is the same despite the context: does one exploit the environment, using current knowledge to acquire a satisfactory solution, or explore other options and potentially find a better answer. An accurate cognitive model must be able to handle this trade-off because of the importance it plays in our lives. We are developing physio-cognitive models to better understand how physiological and cognitive processes interact to mediate decisions to explore or exploit. To accomplish this, we utilize the ACT-R/Φ hybrid architecture (Dancy, 2013; Dancy et al., 2015) and the Project Malmo AI platform (Johnson et al., 2016).

Modelling the Trade-off

ACT-R/Φ

ACT-R/Φ creates a representation of physio-cognitive mediation of behavior by combining the ACT-R theory of cognition, HumMod's physiological model (Hester et al., 2011) and theory from affective neuroscience. This hybrid architecture allows us to model how the management of the exploration versus exploitation trade-off effects the body and mind. Furthermore, the architecture provides a more concrete and tractable method to interact with the model by utilizing concentrations of hormones in the system to influence behavior. Changes in arousal, utility, and decision making can be seen through modifications of hormone concentration and regulation, providing an in depth look at how and why the trade-off is managed.

Model Assumptions

The model makes several assumptions to interact with the task environment. First, the model assumes it is in a diminishing return environment. Second, that cues are present in the environment, which provide information about the task. Lastly, that the agent is striving towards some goal.

Managing the Trade-off

The high-level model manages the exploration exploitation trade-off according to the abstract rules in Figure 1.

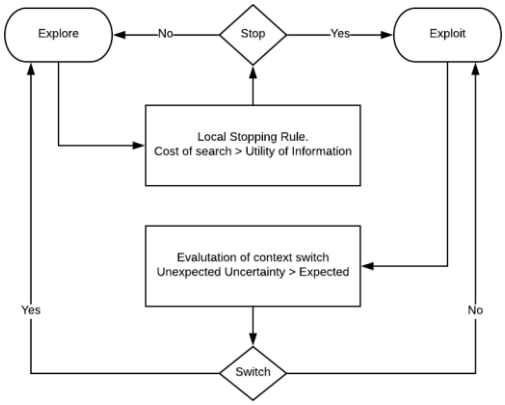

Figure 1. The high-level decisions the model makes to transition between exploring and exploiting.

The model decides to stop exploring via a local stopping rule that integrates the Satisficing Model by Wai-Tat Fu (2006). The model assesses the utility of information, that is, how useful the new information gained from exploring is, against the cost of further exploration. Since the model assumes it's in a diminishing return environment, the cost of searching will be lower at the start and increase as time goes on. Thus, exploration will tend to occur early. Once the cost of search outweighs the benefit of information gained the model will decide how to exploit its knowledge and will continue to exploit until it deems the current method is no longer adequate.

The decision to return to exploration is controlled by assessments of expected and unexpected uncertainty. When unexpected uncertainty is higher than expected, the current method of solving the problem is no longer reliable, thus exploration should start. Yu and Dayan (2005) related these concepts to the neuromodulators acetylcholine (ACh) and norepinephrine (NE). In their formulation, ACh represents expected uncertainty and NE represents unexpected uncertainty. They also developed an equation that relates the concentration of the modulators to the choice to explore (Equation 1).

$$NE > \frac{ACh}{0.5 + ACh} \quad (1)$$

The equation represents low level self-assessment and triggers the transition between exploiting and exploring.

When it is satisfied, subsequent drops in utility and arousal are observed, denoting a loss of faith in the current strategy and need to discover a new one.

The decision making transition is reflected in the model by dynamically modifying the utility noise parameter, also referred to as temperature, in ACT-R. ACT-R selects which production to fire by its utility value. However, those values contain noise. The parameter controls the standard deviation of noise within the system. As noise increases, the probability that the production with the highest reinforcement will be selected decreases. Therefore, the likelihood of the model selecting another, less reinforced, production that satisfies that same scenario increases, leading to exploratory behavior.

As the model receives rewards, the temperature decreases, making production selection more deterministic. This results in the model switching back to the exploiting state. While the model runs, temperature is adjusted, becoming lower when search costs outweigh the value of current information, and larger when self-assessments reveal poor performance.

Testing the Model

We are using a symbolic maze, similar to the one used by Fu and Anderson (2006), to test the model. The structure of the maze is depicted in figure 2.

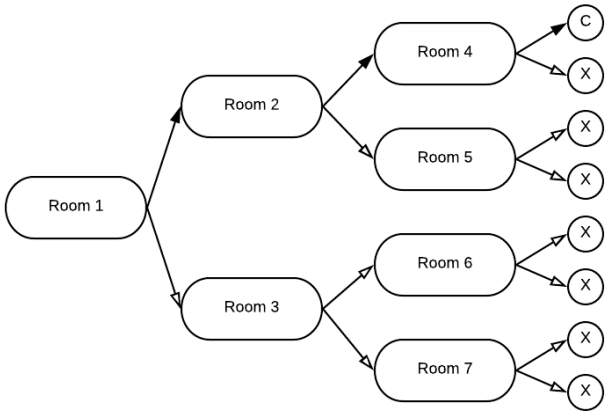

Figure 2. Structure of the symbolic maze. X represents a dead end whereas C depicts the exit. Picture based on an image from Fu and Anderson (2006).

The player is placed in a room and presented with stimuli and a set of options. They move into a different room depending on which option they select. Upon reaching a dead end the player is reset to the point where they diverged from the correct path. Furthermore, the configuration of the room is changed; different stimuli and options are shown upon their return. Thus, the player is only informed of correct stimuli option associations upon completing the maze or reaching a dead end.

The maze is implemented in Microsoft's Project Malmo environment. Project Malmo is a modification to the game Minecraft that allows artificial agents to be tested. The world is represented as a series of semantically defined blocks. This works well with ACT-R based models as the representations in the perceptual modules are semantic attentional chunks. Thus, transforming the *blocks* to *chunks* is straightforward. For our experiment, stimuli is represented by special blocks in a wall. Decisions are made by standing on one of two sections of ore in the floor. After a decision is made, the player or agent is teleported to another room and the experiment continues as previously described.

Another benefit of using Project Malmo is its expandability. The tool can be used to construct varied environments with differing complexities from the same primitives. Using this platform allows us to modify the task to study different aspects of physio-cognitive mediation of human behavior in future work.

Conclusion

Managing the trade-off between exploration and exploitation is a critical part of our everyday lives. Our goal is to develop a model that manages the problem like a human does. We manage the transition from exploration to exploitation by assessing the cost of searching with the utility of information gained. The model handles the inverse transition by low level self-assessments of uncertainty, both expected and unexpected, within the problem. In addition, by using Project Malmo, we have created a useful, modifiable, task environment for future cognitive models. By improving our model to tackle more complex domains within Project Malmo we will be one step closer to developing human-like autonomous artificial agents.

References

Dancy, C.L (2013) ACT-RΦ; A cognitive architecture with physiology and affect. *Biologically Inspired Cognitive Architectures*, 6(1), 40-45.

Dancy, C. L., Ritter, F. E., Berry, K. A., & Klein, L. C. (2015). Using a cognitive architecture with a physiological substrate to represent effects of a psychological stressor on cognition. *Computational and Mathematical Organization Theory, 21*(1), 90-114.

Fu, W. (2007). A Rational-Ecological Approach to the Exploration/Exploitation Trade-offs. In W.D. Gray (ed.), *Integrated Models of Cognitive Systems* (Vol. 1, pp 165-179). New York, NY: OUP.

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.

Hester, R. L., Brown, A. J., Husband, L., Iliescu, R., Pruett, D., Summers, R., & Coleman, T. G. (2011). HumMod: A modeling environment for the simulation of integrative human physiology. *Frontiers in physiology, 2*(12).

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.

Yu, A. J, & Dayan, P. (2005). Uncertainty, Neuromodulation, and Attention. *Neuron, 46*(4), 681-692.