Counterfactual Reasoning as a Key for Explaining Adaptive Behavior in a Changing Environment

Jaehyon Paik[1], Yunfeng Zhang[2] and Peter Pirolli[1]
[1]*Palo Alto Research Center (PARC), 3333 Coyote Hill Rd., Palo Alto, CA 94304*
[2]*Department of Computer and Information Science, University of Oregon*
1202 University of Oregon, Eugene, OR 97403
jpaik@parc.com, zywind@cs.uoregon.edu, pirolli@parc.com

Abstract
It is crucial for animals to detect changes in their surrounding environment, and reinforcement learning is one of the well-known processes to explain the change detection behavior. However, reinforcement learning itself cannot fully explain rapid, relatively immediate changes in strategy in response to abrupt environment changes. A previous model employed reinforcement learning and counterfactual reasoning to explain adaptive behavior observed in a changing market simulation environment. In this paper, we used the same model mechanisms to simulate data from two additional tasks that require participants, who played the role of intelligence analysts, to detect the changes of a computer-controlled adversary's tactics based on intelligence evidence and feedback. The results show that our model captures participants' adaptive behavior accurately, which further supports our previous conclusion that counterfactual reasoning is a missing piece for explaining adaptive behavior in a changing environment.

Keywords: detecting changes, reinforcement learning, counterfactual reasoning, ACT-R cognitive model.

1. Introduction

It is crucial for animals and humans to detect changes in their surrounding environment, which can happen either gradually or drastically. Animals' survival depends on how well they adapt to these environmental changes, and learning is perhaps the most powerful ability that animals posses to cope with these changes.

Studies on change detection argue that animals use reinforcement learning (Sutton & Barto, 1998) to detect environmental changes (Behrens, Woolrich, Walton, & Rushworth, 2007; Pearson, Heilbronner, Barack, Hayden, & Platt, 2011). Reinforcement learning is theoretically similar to linear operators, which have been shown, based on optimal foraging theory, to track the changes of a hidden environmental variable with probabilistic observations (McNamara & Houston, 1987). Several behavioral and neuroimaging studies showed that people seem to use reinforcement learning to detect

changes, and their performance in the tasks approaches the performance of an ideal observer (Behrens et al., 2007; Nassar, Wilson, Heasly, & Gold, 2010).

Although reinforcement learning plays a key role in detecting changes, it alone cannot fully explain how some animals quickly switch to different task strategies because the error-learning rule for reinforcement learning can only explain a gradual transition between strategies in response to abrupt changes (Pearson et al., 2011). In a previous study (Zhang, Paik, & Pirolli, in press), we showed that counterfactual reasoning, a cognitive strategy that considers what would happen if an option different from the selected option is carried out, might be the key to explaining rapid detection of environmental changes. In that study, we conducted an experiment and asked participants to try to earn as much money as possible by investing in a virtual market that periodically switches between a bear and a bull state. It was found that participants were able to make nearly optimal decisions about when to invest in the market and when to skip the investment opportunity to avoid likely losses. Furthermore, we found that a model that incorporated reinforcement learning and counterfactual reasoning was able to explain the behavioral data, whereas a model that only implemented reinforcement learning could not. In this study, we follow the same approach and develop an ACT-R model for two additional tasks to provide further evidence to support our hypothesis that counterfactual reasoning is a missing piece for explaining change detection behaviors.

2. Change Detection Tasks

The IARPA ICArUS program developed a series of five tasks which, collectively, are called TACTICS. These tasks simulate some common intelligence analysis missions, in which the analysts need to predict and sometimes counteract an opponent's actions. TACTICS is the successor to the ICArUS challenge tasks (Lebiere, Pirolli, Thomson, Paik, Rutledge-Taylor, Staszewski, & Anderson, 2013), and both projects were designed to drive the development of integrated neurocognitive models of sensemaking. In this paper, we report the results of two tasks, Missions 4 and 5, which required participants to detect the changes of a simulated opponent's tactic based on intelligence evidence and feedbacks.

In TACTICS, a participant (Blue defense) operates against a computer agent (Red offense) over a series of trials in an area of interest using intelligence data depicted on a Geographic Information System display as can be seen in Figure 1. Each trial involves a particular location (indicated by the blue dot and the green circle) in Blue's territory (outlined by the blue lines). Two pieces of intelligence information are given for each trial: (a) OSINT (open source intelligence, P), which indicates the probability that Blue will defeat Red (Blue's vulnerability); and (b) IMINT (imagery intelligence, U), which indicates the utility/payoff at stake in a showdown (Opportunity). Based on these two INTs, Blue can estimate the Red attack probability and then select either (a) *Divert*, to avoid a possible Red attack or (b) *~Divert*, to counter the Red attack. If Blue wins a showdown (Red attack and Blue ~Divert), then Blue earns U units of utility. If Red wins, then blue loses U utility. Therefore, Blue needs to consider its vulnerability (P) and opportunity (U) to make a decision for its actions. Blue loses 1 unit of utility when it selects Divert and Red does not attack. Blue neither wins nor loses in the other two cases (Red attacks and Blue selects Divert, or Red does not attack, and Blue selects ~Divert).

In Missions 4 and 5, Red tactics change at some point in time from one particular attack probability distribution to another distribution (more details will be discussed below). To complete the missions successfully, Blue needs to estimate the Red attack probability based on its hypothesis about Red tactics and detect when the tactics change.

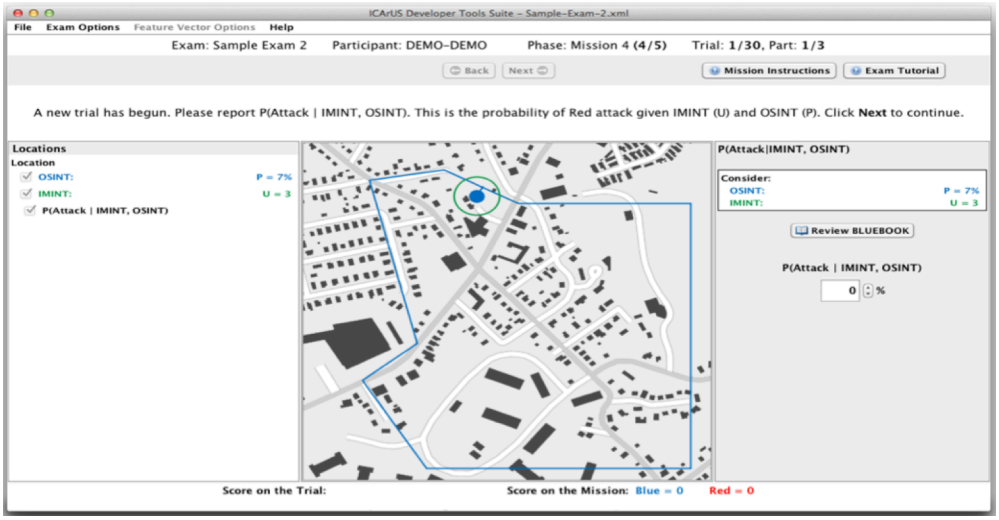
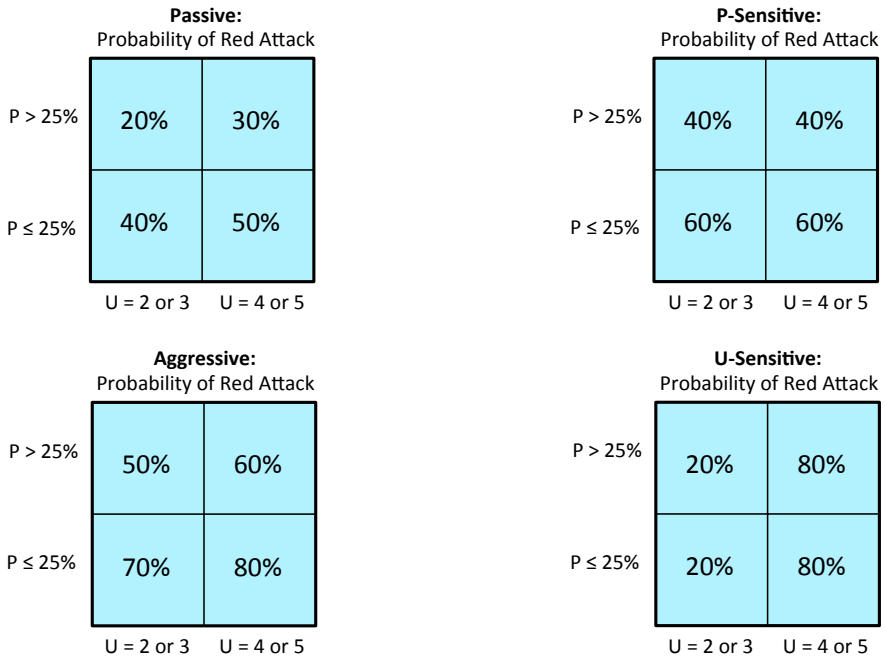


Figure 1: A screenshot of the display of the TACTICS task. The left column displays the values of intelligence information (OSINT and IMINT) for a particular location on the map (middle column). The middle column provides geospatial information about an interest location, Blue's territory, and the density of buildings. The right column provides an input box that participants can enter the estimated Red attack probability based on two INTs.

In Mission 4, Red's tactic begins as either *Passive* or *Aggressive*, and later changes to the other tactic at some point during the 30 trials of a mission. As can be seen in Figure 2-a, the Passive tactic generally has lower attack probabilities than the Aggressive tactic with the same ranges of OSINT and IMINT, which indicates that the Red's attack frequency is the key factor to estimate Red's tactic. In all but the first trial, participants are asked to estimate the tactic that Red is currently using by assigning probabilities for each tactic. Then, OSINT and IMINT values of the current trial are provided along with propensity tables (See Figure 2-a) for both tactics. Based on this information, participants are asked to estimate Red's attack propensity, and to select "Divert" or "~Divert". Finally, the Red action is displayed with the resulting payoff value. Mission 5 is almost the same as Mission 4 except that the number of trials in Mission 5 is 40 rather than 30, and that the Red's possible tactics are *P-Sensitive* and *U-Sensitive* rather than *Passive* and *Aggressive*. As can be seen in Figure 2-b, the P-Sensitive tactic is susceptible to the amount of OSINT value. For example, if P is less than 25%, then the likelihood of the Red attack is 60% when the Red uses the P-Sensitive tactic. However, when the Red uses the U-Sensitive tactic, it is more susceptible to the amount of IMINT value with larger U has larger attack probability.

In Mission 4, Red's tactics can be easily inferred from the frequency of past attacks because the Red attack probability for the passive tactic is always less than 50% and for the aggressive tactic, always greater than 50%. However, in Mission 5, Red tactics cannot be easily inferred only from the attack frequency of past attacks, and participants need to pay attention to the values of P and U in subsets of past attacks.



(a) (b)

Figure 2. Probabilities of Red attack based on Red tactics in Mission 4 (a) and Mission 5 (b).

3 ACT-R Cognitive Architecture and Model

We developed an ACT-R model of change detection tasks based on reinforcement learning and counterfactual reasoning, which are the same approaches that we developed in another change detection task (Zhang et al., in press). ACT-R (Anderson, Bothell, Byrne, Douglass, Lebiere, & Qin, 2004; Anderson & Lebiere, 1998) is a cognitive architecture to simulate and understand human cognition. It has many built-in modules, such as declarative and procedural modules, to represent the flow of information in human mind. Recent study (Anderson et al., 2004) showed that each of the modules in the ACT-R architecture could be mapped to a particular brain region using fMRI data.

The procedural module of ACT-R coordinates the flow of information learned by reinforcement learning. In the procedural module, each tactic in our tasks can be represented as a production rule, which is an if-then condition-action pair. When the conditions are met, the actions are executed, and the productions can be rewarded or penalized by the reinforcement learning mechanism. In ACT-R, each production rule has a utility value, which roughly corresponds to how likely the production rule leads to the successful completion of the task. In every 50-ms cognitive cycle, ACT-R executes one of the production rules whose conditions are matched, and the probability that a matched rule will be selected is an increasing function of its production utility:

$$Probability(i) = \frac{e^{U_i/\sqrt{2}s}}{\sum_j e^{U_j/\sqrt{2}s}}$$

where U_i is the utility of the production rule i, s is the utility noise, which is set with the parameter :egs in the ACT-R architecture, and the denominator is a summation over all production rules that have matched conditions.

When a task goal is reached (or fail) and a reward (or penalty) is triggered, the reward (penalty) is propagated back through the firing chain of the productions rules so that the utility of the previously fired rules can be updated according to the following equation:

$$U_n = U_{n-1} + \alpha(R_n - U_{n-1})$$

where U_{n-1} is the utility of the production rule before the update, U_n is the utility after the update, R_n is the reward, and α is the learning rate. The equation is based on a *temporal difference* learning rule (Sutton & Barto, 1990), which is the extension of Rescorla-Wagner model (1972) to explain the timing of different events.

In addition to reinforcement learning, our model also incorporated counterfactual reasoning, which is a cognitive process that enables the model to evaluate the alternative choice and makes the model to capture changes more rapidly. That is, the current hypothesis (tactic) only gets rewarded (or penalized) according to the reinforcement learning mechanism, however, the alternative hypothesis is also evaluated and updated (rewarded or penalized) by the counterfactual reasoning mechanism. The reinforcement learning alone cannot predict the rapid, relatively immediate changes in tactics in response to environmental changes. By incorporating counterfactual reasoning, our model could consider both tactics and update their utilities in each trial, which helps the model detect changes more rapidly.

Figure 3 illustrates the processes of the change detection model. There are two production rules, which stand for two tactics (*Aggressive* or *Passive* in Mission 4, *P-Sensitive* and *U-Sensitive* in Mission 5) and are competing each other during the task. The model starts with selecting a particular tactic, and based on OSINT and IMINT values the model retrieves the propensity of each tactic from the propensity tables (see Figure 2), which are the rewards and penalties that the model uses during the

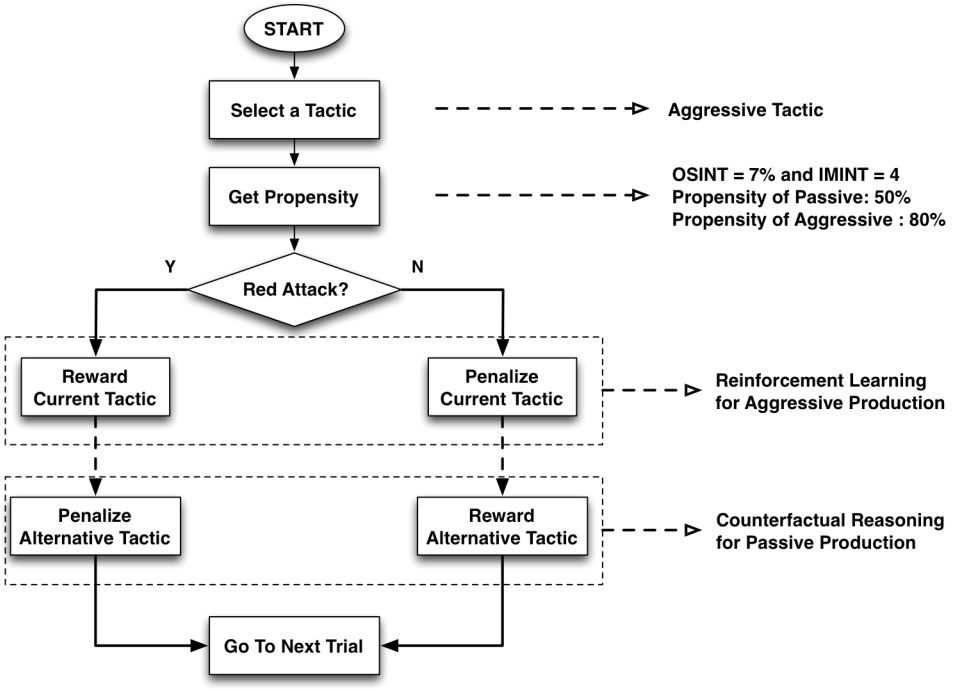


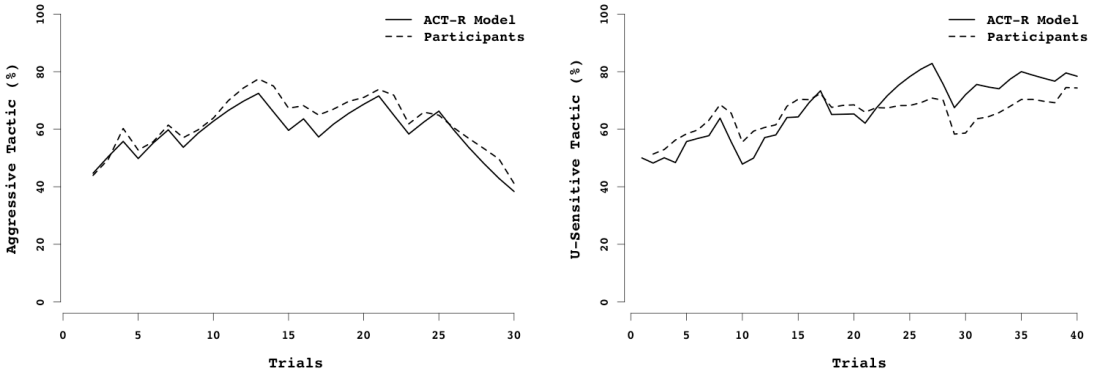
Figure 3: A flow chart showing how the model incorporates reinforcement learning and counterfactual reasoning in a particular trial.

next processes, reinforcement learning and counterfactual reasoning. For example, if OSINT is 7% and IMINT is 4 in a particular trial, then Red attack probability (propensity) for the Passive tactic is 50%, and for the Aggressive tactic is 80%. If the model's current hypothesis for the tactic is aggressive and Red actually attacks in that particular trial, then the *Aggressive* production rule gets 100 * 80% as a reward. If Red does not attack, then the *Aggressive* production rule gets 100 * (1-80%) as a reward. These are the basic mechanism of reinforcement learning in ACT-R. By the counterfactual reasoning approach, which is the same reinforcement learning mechanism for the alternative hypothesis, Passive, our model updates the *Passive* production rule according to the value of Passive propensity table and Red decision. That is, the *Passive* production gets 100 * 50% if Red attacks, and 100 * (1-50%) if Red does not attack.

4 Human Data and ACT-R Model Prediction

A total of thirty participants who were graduate and undergraduate of the Pennsylvania State University were recruited by the MITRE Corporation, and the participants performed Missions 1 through 5 with resting periods between the missions. We analyzed the participants' average estimation of each tactic, Passive and Aggressive for Mission 4, and P-Sensitive and U-Sensitive for Mission 5, at each trial, and compared participants' estimation with our ACT-R model prediction.

Figure 4 (a) shows the ACT-R model's predictions on the Aggressive tactic probability at each trial and participants' average estimation for the same tactic in Mission 4. The fit between the ACT-R model estimation and average participants' estimation is $R^2 = 0.92$, which indicates the ACT-R model captures participants' estimation of tactics very accurately. Figure 4 (b) shows the results of the ACT-R model and participants' estimation for the U-Sensitive tactic in Mission 5. The fit between the ACT-R model and participants is $R^2 = 0.61$.



(a) (b)

Figure 4. (a) The results of our ACT-R model and participants' estimation of probability for the Aggressive tactic in Mission 4 and (b) the results of our ACT-R model and participants' estimation of probability for the U-Sensitive tactic in Mission 5.

The opponent's tactic starts with Aggressive and changes to Passive at trial 21 in Mission 4. It took 9 trials on average for participants to detect changes in tactic (probability estimation for Aggressive tactic < 50%), and it took 8 trials for the ACT-R model to detect change in Mission 4.

In Mission 5, the opponent's tactic starts with P-Sensitive and changes to U-Sensitive at trial 9. It is not clear that participants ever detected the change in this mission, because their estimated

probability for the U-Sensitive tactic fluctuated across the trials. Interestingly, the ACT-R model also captures this estimated probability fluctuation in this mission, although model's probability estimation is deviated from participants' estimation after 21st trial.

5 Conclusions and Discussion

Our previous study showed that people could detect changes in a stochastic environment, and the cognitive model that incorporated reinforcement learning and counterfactual reasoning explains this adaptive behavior accurately. In this paper, we describe two additional tasks that require participants to detect the changes of opponent's tactic based on intelligence evidences and feedbacks, and a cognitive model that predict participants' adaptive behavior based on reinforcement learning and counterfactual reasoning.

It took 9 trials on average for participants to detect the tactic change after the actual change happens in Mission 4, and it was not clear that participants detected changes in Mission 5, because their probability estimation for each tactic fluctuated across the trials.

Our cognitive model that incorporates reinforcement learning and counterfactual reasoning seems to accurately account for participants' estimation in both missions. However, there were some deviations for estimation between the model and human data after the 21st trail in Mission 5. We investigated those trials to find out these deviations, and found that the U-Sensitive tactic was rewarded 80 from 21st to 26th consecutively while the P-Sensitive tactic was reward 40 in the same trials based on reinforcement learning and counterfactual reasoning mechanisms, which make the estimation of U-Sensitive tactic increases sharply. However, participants' average probability estimations at those trials changed less substantially. We can assume that participants might be exhausted during performing the entire missions or participants' perception for the amount of the probability might differ from the model's estimation. The different characteristics for each mission might be the reason. That is, in Mission 4, the opponent's tactic can be easily inferred from the total frequency of past attacks, however, in Mission 5, participants' needed to consider not only the frequency of past attacks, but also the values of P (OSINT) and U (IMINT).

Although there are some deviations between our model and human in Mission 5, the model still captures the main trends in the human data for both missions. Furthermore, our previous study also captures the adaptive behavior in a changing environment using reinforcement learning and counterfactual reasoning. From those two studies, we suggest that counterfactual reasoning might be a hidden mechanism, working with reinforcement learning for explaining adaptive behavior in rapidly changing environments.

Acknowledgement

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract number D10PC20021. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained hereon are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*(4), 1036-1060.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature neuroscience, 10*(9), 1214-1221.

Lebiere, C., Pirolli, P., Thomson, R., Paik, J., Rutledge-Taylor, M., Staszewski, J., & Anderson, J. R. (2013). A Functional Model of Sensemaking in a Neurocognitive Architecture. *Computational Intelligence and Neuroscience*.

McNamara, J. M., & Houston, A. I. (1987). Memory and the efficient use of information. *Journal of Theoretical Biology, 125*(4), 385-395.

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *The Journal of Neuroscience, 30*(37), 12366-12378.

Pearson, J. M., Heilbronner, S. R., Barack, D. L., Hayden, B. Y., & Platt, M. L. (2011). Posterior cingulate cortex: adapting behavior to a changing world. *Trends in cognitive sciences, 15*(4), 143-151.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory, 2*, 64-99.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of pavlovian reinforcement.

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning*: MIT Press.

Zhang, Y., Paik, J., & Pirolli, P. (in press). *Reinforcement Learning and Counterfactual Reasoning Explain Adaptive Behavior in a Changing Environment.* Paper will be presented at the 36th Annual Conference of the Cognitive Science Society, Quebec City, Canada.