Reinforcement Learning and Counterfactual Reasoning Explain Adaptive Behavior in a Changing Environment

Yunfeng Zhang (zywind@cs.uoregon.edu)
Department of Computer and Information Science, University of Oregon
1202 University of Oregon, Eugene, OR 97403

Jaehyon Paik (jpaik@parc.com) and Peter Pirolli (pirolli@parc.com)
Palo Alto Research Center
3333 Coyote Hill Rd., Palo Alto, CA 94304

Abstract

Animals routinely adapt to changes in the environment in order to survive. Though reinforcement learning may play a role in such adaption, it is not clear that it is the only mechanism involved, as it is not well suited to producing rapid, relatively immediate changes in strategy in response to environmental changes. We explored the possible adaptive mechanisms underlying in a cognitive model of human behavior in a change detection experiment. Besides reinforcement learning, the model incorporates counterfactual reasoning to help learn the utility of different task strategies under different environmental conditions. The results show that the model can accurately explain human data and that counterfactual reasoning is key to reproducing the various effects observed in this change detection paradigm.

Keywords: change detection, reinforcement learning, counterfactual reasoning, cognitive modeling.

Introduction

Detecting changes in the natural environment is often vital for an organism's survival. Animals routinely experience environmental changes across days and seasons, and sometimes more sudden and drastic changes such as flood and drought. Evolution has equipped organisms with many abilities to detect such changes, and learning is perhaps the most powerful one of them. Studying change detection, a problem that learning is possibly originally evolved for, may shed light on the capabilities and limitations of learning.

Rational analyses of change detection have been developed based on optimal foraging theories (e.g., McNamara and Houston, 1987; Stephens, 1987). Stephens (1987) derived the optimal foraging strategies for a simplified, hypothetical environment that contains a variable food energy resource that periodically switches between a poor and a rich state, and a stable food energy resource that provides a medium amount of energy. It is found that to maximize the intake of food energy, there is an optimal frequency for how often the variable resource should be sampled (to detect its rich state). This analysis suggests that to survive in the natural world, animals need to actively explore the environment and perhaps need to do so in a particular rate to maximize the benefit and minimize the cost of explorations. But how does animals learn when to explore, and what mechanisms drive them to explore rather than to stay in a stable habitat?

Past research suggests that animals may use reinforcement learning to detect environmental changes (Behrens et al., 2007; Pearson et al., 2011). Reinforcement learning was shown to be a biologically plausible learning mechanism (Holroyd and Coles, 2002) and it is very similar to linear operators derived in optimal foraging theory to track the changes of a hidden environmental variable with probabilistic observations (McNamara and Houston, 1987). Several behavioral and neuroimaging studies (Behrens et al., 2007; Nassar et al., 2010) showed that people seem to use reinforcement learning to detect changes, and their performance in these tasks approaches the performance of an ideal observer.

Despite its dominance in the discussion of change detection, reinforcement learning alone cannot fully explain how some animals often quickly switch to drastically different task strategies, because its error-learning rule suggests a gradual transition of behaviors in response to changes (Pearson et al., 2011). For example, reinforcement learning cannot easily explain how monkeys do not just try to jump higher to reach a bunch of hanging bananas, but know to use chairs and sticks. Such strategies cannot result from gradual updates of a single strategy, rather, they are likely a result of evaluating a wide array of different options.

This research proposes that counterfactual reasoning is a missing piece in this theoretical framework for explaining change detection behaviors. Counterfactual reasoning captures the process in which humans think about potential or imaginary events and consequences that are alternatives to what has occurred. This gives humans abilities to learn the utility of a task strategy without actually applying it. Neuroimaing studies (e.g., Coricelli et al, 2005) show that such processes indeed exist and they seem to play a key role in decision making. Nevertheless, counterfactual reasoning is somewhat overlooked as a plausible explanation for change detection behaviors.

This paper presents the behavioral data collected from a stochastic change detection task, compares the human data with the predictions of an ACT-R cognitive model that incorporates reinforcement learning and counterfactual reasoning, and compares models with and without counterfactual reasoning to demonstrate the importance of counterfactual reasoning in explaining human change detection performance.

Experimental Paradigm

The change detection experiment presented here is designed as an investment game, in which there is a virtual market that the participant can invest virtual chips in. The market alternates between the *bear state*, in which the participant is likely (> 50%) to lose the investment, and the *bull state*, in which the participant is likely (> 50%) to profit. The change of the market state occurs at a small, constant probability per turn. The market state is not directly observable by the participants, but has to be inferred from the investment outcomes (profiting or losing) of the recent trials. In essence, the virtual market is designed as a hidden Markov process to mimic the natural environment in which the underlying states, such as the amount of food in a habitat, are not directly observable, but are often similar to the states of the recent past.

Participants Forty-eight participants (26 females; mean age = 36.71 years, range 21–62 years) were recruited on the Amazon Mechanical Turk website. Each participant received a base compensation of $3 and up to a $4 bonus for completing the 30 min long experiment. The bonus that participants received depended on their task performance.

Apparatus and Materials In each trial, two buttons labeled "pass" and "10" were presented on the screen. Clicking "pass" would skip the investment opportunity, while clicking "10" would invest 10 chips to the market. If the participant chose to invest, he or she would either win 15 chips or lose the 10-chip investment. This investment outcome, as well as the participant's total number of chips, were immediately shown to the participant after each trial. If the participant finished a trial within 5 seconds (indicated by a count-down timer on the display), a reward of 0.05 cents would be added as a bonus.

Design Three factors were manipulated. The first was the discriminability between the bull market and the bear market. **Error! Reference source not found.** shows the probability of profiting and losing in the two discriminability conditions tested in the experiment. As can be seen from the table, the profiting probability of the bull and the bear markets were set to be more similar in the low discriminability condition than those in the high discriminability condition, and hence it was harder to distinguish the two market states in the low discriminability condition. Manipulating this factor helped us examine how the reliability of observation might affect peoples' ability to infer the underlying environmental states.

The second factor of the experiment was the probability of a market-state change in each trial. Again, two levels were tested, one with 5% change probability and the other with 15% change probability. This factor examined how well people adapt to the volatility of the environment.

Table 1: The probability of profiting and losing of the bear and the bull markets in the low and the high discriminability conditions.

	Low Discriminability		High Discriminability	
	Profiting	Losing	Profiting	Losing
Bull	70%	30%	10%	90%
Bear	30%	70%	90%	10%

The third factor of the experiment was whether to provide information about the outcome of the market when "pass" was selected. In the no-feedback-for-pass condition, the participant needed to guess what was happening in the market if pass was selected, based on the past experience such as how long the bear market generally lasted. This condition simulated an environment in which one can only acquire information about the choice they made. We expected that participants would perform worse in the no-feedback-for-pass condition than in the has-feedback-for-pass condition.

The discriminability factor was a within-subject variable, and the change-frequency and the feedback factors were between-subject variables balanced across the 48 participants. In other words, each participant did both the low discriminability and the high discriminability conditions, but experienced only one change frequency and one feedback condition.

Procedure The participant clicked a link provided on an Amazon Mechanical Turk webpage to navigate to the experiment website. Before doing the experiment, the participant needed to accept a consent form, fill out a demographic survey, and complete a risk propensity scale (see Meertens and Lion, 2008). The experiment instructions included that the market switches between a bull and a bear state at a constant probability per trial, but no concrete parameters such as the profiting probabilities were shown to the participants.

Each participant completed two low-discriminability blocks and two high-discriminability blocks, with the running order randomized and balanced across participants. The participant was informed about the market discriminability before each block. In each block, the participant started with 100 chips, and underwent 150 trials. The performance feedback, including the number of chips earned and how much bonus the chips translated to, was provided after each block and at the end of the experiment.

Experimental Results

Figure 1 shows the overall task performance across the different experimental conditions, measured as the average number of chips earned in each block. Participants earned more chips in the high-discriminability condition than in the low-discriminability condition ($z = -11.2$, $p < .001$)[1], as is shown in the graph that the bars in the left column are taller than the bars in the right column. Participants also earned more chips in the has-feedback condition than in the no-feedback condition ($z = -2.45$, $p < .001$), as is shown in the

[1] Multiple comparisons were done using general linear hypotheses tests on a linear mixed-effects model. Main effects and effect sizes were obtained using a repeated measure ANCOVA, with the covariate being the number of profitable trials in a block.

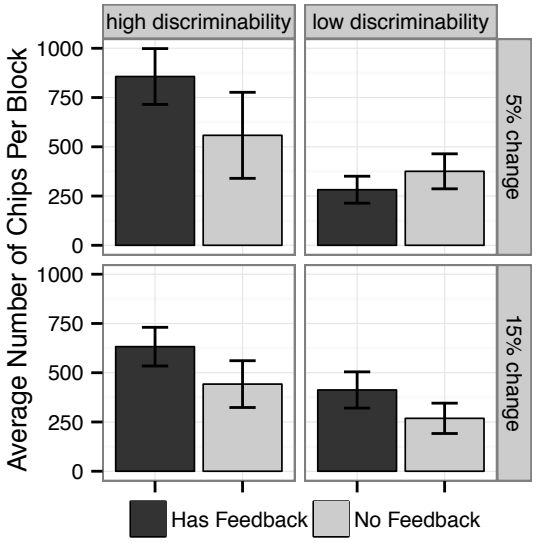

Figure 1: Average number of chips earned per block in the different experimental conditions. Error bars show 95% confidence intervals of the mean.

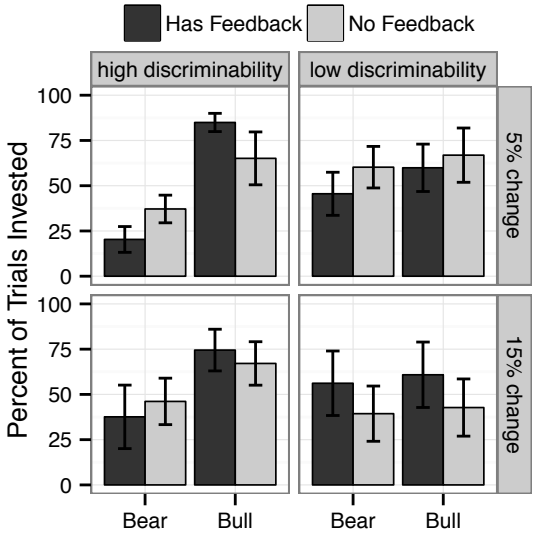

Figure 2: Average percentage of trials the participants invested in when the market was in the bear and the bull states in the different experimental conditions. Only the last 100 trials of each block were used for this graph.

graph that within each panel, the dark gray bar is usually taller than the light gray bar. Market discriminability and the feedback condition had the largest effect on task performance (for discriminability, $F(1, 44) = 55.5$, $p < .001$, $\eta_G^2 = .888$; for feedback, $F(1, 44) = 7.42$, $p = .009$, $\eta_G^2 = .12$), whereas change frequency did not have a significant main effect, $F(1, 44) = 3.00$, $p = .09$, $\eta_G^2 = .056$.

Figure 2 reveals participants' investment strategies and shows how these strategies are heavily influenced by the market discriminability. The investment percentages in the graph were calculated using the last 100 trials of each block because at the beginning of each block participants were likely still exploring the task parameters, and the behaviors during the first 50 trials probably cannot represent the stabilized behavior. As can be seen from the graph, in the high-discriminability conditions (left column), the investment percentages of the bear market are very different from those of the bull market, particularly in the top left panel. This result suggests that the participants could somewhat accurately infer the market state and use that information to avoid investing in the bear market and at the same time, exploit the bull market. In the low-discriminability condition, however, the investment percentages are about the same across the bear and bull markets. This suggests that the participants could not identify the market state and thus applied the same strategy all along, which no doubt contributed to the poor performance in the low-discriminability conditions.

Figure 2 also shows that in the no-feedback condition, in which the market outcome was only provided if "invest" is selected, participants were less able to detect the underlying changes of the market. This can be seen in the left two graphs in Figure 2 (high discriminability) where the difference between the no-feedback conditions is less than the difference between the has-feedback conditions. Furthermore, it seems that when there was no feedback for the pass option, the participants were more conservative in investing in the market, as can be seen that in the bottom right panel the no-feedback conditions has smaller investment percentages than the has-feedback conditions.

The above results suggest that the participants' strategies may be rational to some extent because they tried to maximize their pay in some conditions, but their ability to infer the market state from the probabilistic observations is perhaps limited by memory and learning mechanisms. The next section presents a cognitive model that tries to reproduce this bounded rationality using reinforcement learning and a counterfactual reasoning strategy.

The Change Detection Model

The model presented here is implemented using the ACT-R cognitive architecture (Anderson et al., 2004). ACT-R has many built-in constructs that directly support the modeling of this task. Particularly, it has a powerful production system that learns by reinforcement learning. In a production system, task strategies are written as production rules, which are IF-THEN statements that execute certain actions (the THEN part) when the conditions are met (the IF part). In ACT-R's production system, each production rule can also be assigned a utility value, which roughly corresponds to how likely this 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:

$$Probability(i) = \frac{e^{U_i/\sqrt{2}s}}{\sum_j e^{U_j/\sqrt{2}s}} \quad (1)$$

where U_i is the utility of the production rule i, s is a free parameter, and the denominator is a summation over all production rules whose conditions are matched. s is also referred to as the utility noise parameter, because as s increases the probability that a production rule will fire depends less on its utility and more on the random chance.

When a task goal is reached (or fails) and a reward (or penalty) is triggered, the reward (penalty) is propagated back through the firing chain of the production rules so that the utility of the previously fired rules can all be updated accordingly by the following equation:

$$U_n = U_{n-1} + \alpha(R_n - U_{n-1}) \quad (2)$$

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 production selection equation and the utility updating equation are essentially the same as the ones used in some ideal observer models (Behrens et al., 2007; Nassar et al., 2010) with the exception that the learning parameter α in ACT-R is set by the analyst, as opposed to be learned on a trial-by-trial basis.

Figure 3 illustrates the task strategy of the model. At the beginning of the trial, the model executes one of the two production rules, *assume-bull* and *assume-bear*, based on Equation 1. If *assume-bull* fires, the rule *invest* will ensue because it is rational to capitalize on the bull market. Then just like the experimental design, if the market returns a profit, a reward of 15 will be delivered and the utility of *assume-bull* will be updated using Equation 2; if the market returns a loss, a penalty of 10 ($R_n = -10$) is delivered. If *assume-bear* fires, the rule *pass* will be fired next, and a reward of 0 will be delivered just like how the participant would neither win nor lose when selecting pass.

Note that the model does not explicitly track the environmental parameters such as the profiting probabilities of the bull and bear markets, which might hinder its ability in making correct investment decisions. This deficiency is somewhat compensated by the utility updating equation that automatically incorporates the frequency in which the reward and penalty occur. When the market state is stable, the utility of *assume-bull* and *assume-bear* should, over time, tend to the expected return of the bull and bear markets, and the production selection based on these utilities should lead to good investment decisions.

The model tracks the change of the market state by heavily weighting the experience of the recent trials when updating the utility of *assume-bull* and *assume-bear*. The learning parameter α is set to 0.5 to give equal weights to the recent experience and to the last utility estimation, which enables the production utility to quickly respond to the change of the market. For example, considers how the model would detect the change from the bull to the bear state. Initially, the model will continue firing *assume-bull* because this rule accumulated high utility from winning in the bull market. But as losing becomes more frequent after the market changes to the bear state, the utility of *assume-bull* is penalized and quickly drops down to below zero, at

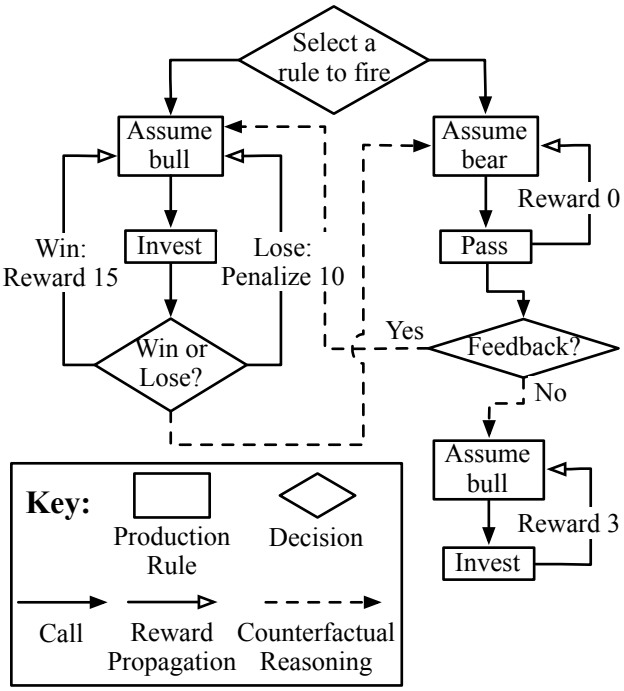

Figure 3: A flow chart showing how the model performs a trial of the experiment. Each trial only goes through one of the dashed lines once to complete the counterfactual reasoning (CR) process. In CR, the production utilities are updated the same way as in a regular learning process.

which point *assume-bear* is fired because its utility (which stays at zero) is now larger than the utility of *assume-bull*.

To detect the change from the bear market to the bull market, however, requires counterfactual reasoning, which evaluates what would happen if the non-selected choice was selected given the newly acquired information about the environment. For the proposed model, if there is no counterfactual reasoning, then once *assume-bear* is selected, the model is likely to be trapped in an assume-bear state, especially when the production noise parameter s is set low. This is because when *assume-bear* fires, *assume-bull*'s production utility is likely below zero. To reset its utility, it needs to be fired, but because *assume-bear*'s utility is higher, it does not have a chance to fire. With counterfactual reasoning, the model temporarily disables *assume-bear* so that *assume-bull* has no competition and can be fired. This way, the model can appropriately update *assume-bull*'s utility when the market changes to a bull state, which then allows the detection of the change.

The counterfactual reasoning processes used by the model are indicated in Figure 3 by the dashed-line connections. As can be seen, after evaluating the made choice, the model continues to another path to evaluate and update the utility of the alternative choice. Note that the lower-right corner of the graph specifies what to do when the current condition does not provide feedback about the market outcome for the pass option (No Feedback condition). In this situation, because the model does not know what would occur if it

invested in the market, it needs to estimate a reward for *assume-bull*. We explored a few settings for this reward parameter and set it at 3 in the final model so that the model generates streaks of pass choices (which are eventually interrupted as *assume-bull*'s utility surpasses *assume-bear* through counterfactual reasoning) that are about as long as those observed in the empirical data.

Overall, the model is a straightforward combination of reinforcement learning and counterfactual evaluation. As will be shown in the following section, though the model does not perform as well as an optimal Bayesian model in terms of the number of chips earned, it does seem to fit the human data.

Model Results

The model was run on all 28,800 trials that the participants performed. To examine whether the model and the participants achieved optimal performance, a Bayesian optimal solution was developed. For every trial, this solution computes the posterior probability of the bull and bear markets given the market outcome and the prior probability of the two markets (which are computed from the previous trial using the same procedure). It then calculates the expected return of investing, and if the return is smaller than zero, pass will be selected, otherwise, investing will be selected. Unlike our human participants, this Bayesian model has knowledge of the underlying market profitabilities (70%/30% or 90%/10%) and underlying change probabilities (5% or 15%), and can thus make optimal decisions. The human data, the model predictions, and the optimal solutions are compared below.

Figure 4 shows the investment percentages across the three data sets. As can be seen, the model (light gray) match the human data (dark gray) very well in almost all conditions except in the No-Feedback group's top-left and bottom-right panel. Similar to the participants, in the high discriminability condition, the model was able to capitalize on the bull market and avoid investing in the bear market, whereas in the low discriminability condition, the model invested at similar percentages across the two markets. In the conditions in which the model does not match the data well (No-Feedback condition's top-left and bottom-right panel), the model invested more aggressively than the participants. Further examination of the payoff data shows that the model in fact earned more chips than the participants (by 0.5 chips per turn) in these conditions, which suggests that the model's strategy—always performing counterfactual reasoning—is a "good enough" strategy, and perhaps the reason that participants did worse is because they did not always use counterfactual reasoning.

The model matched the human data well even in conditions in which the participants' strategy deviated from the optimal solution. It can be seen from Figure 4 that the optimal solution matches the participants' and model's strategies in almost all conditions except in the low-discriminability and 15%-change condition. This condition is the most difficult condition of the experiment, and indeed

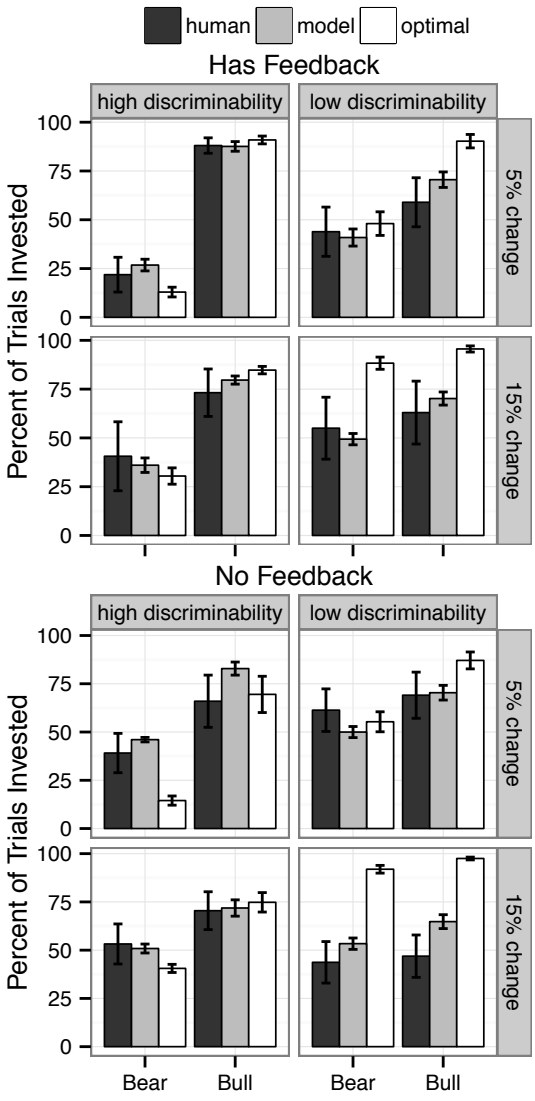

Figure 4: Average percentage of trials invested by the participants, the model, and the optimal Bayesian solution. Only the last 100 trials of each block were used in this analysis.

even the optimal solution could not distinguish the bear and bull markets and had to adopt a uniform investment percentage across the two markets. Unlike the participants and the model, however, the optimal solution invested very aggressively, almost at 100%, in both markets, whereas the model and the participants only invested in about 50% of the trials. The reason that the model could reproduce the participants' conservative strategy is perhaps that when the environment is volatile, the model never had the chance to learn the expected return of the bull market because whenever the model starts investing, the frequent losses soon leads to a switch to the pass behavior. The utility of assume-bull thus remained low most of the time, which resulted in a conservative behavior.

Figure 5 illustrates how counterfactual reasoning (CR) is an indispensable component of the model for explaining the human data. The *y*-axis shows the average absolute

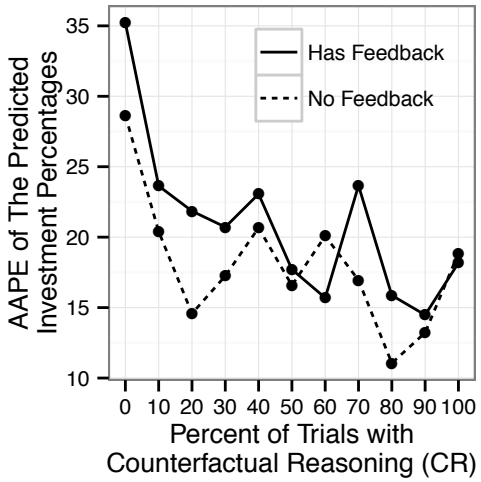

Figure 5: Average absolute percentage error (AAPE) between the predicted investment percentages and the observed percentages, for the 11 models that utilize counterfactual reasoning (CR) at different rates.

percentage error (AAPE) between a model's predicted investment percentages and the observed percentages. In this analysis, we created 11 versions of the model that perform CR at different frequencies, ranging from 0% of the trials to 100% of the trials. It can be seen that if the model never uses CR (0%), its predictions are about 30% to 35% away from the observed investment percentages. As the model utilizes CR more frequently, the predictions become closer to the human data. The best fit is reached at somewhere between 80% CR and 90% CR, which suggests that perhaps participants did CR most but not all of the time.

Discussion and Conclusions

Our experimental results show that people can detect changes in a stochastic environment in which the observations are only imperfect indicators of the environment's underlying state. When the observations can be used to somewhat reliably identify the hidden states, the participants' performance approach optimal. When the observations do not reliably identify the hidden states, participants seem to show loss aversion and to adopt a conservative strategy to avoid risks.

A cognitive model that uses reinforcement learning and counterfactual reasoning seems to accurately account for participants' performance, be it optimal or suboptimal. The fact that the model has very few free parameters and yet it can still predict the trends in the human data across a variety of conditions strongly suggests that reinforcement learning and counterfactual reasoning might be the main mechanisms behind decision making in such changing environment. Particularly, that the model reproduces participants' tendency of loss aversion in the most volatile condition suggests that perhaps loss aversion is simply a byproduct of applying reinforcement learning in a very unpredictable environment.

A model sensitivity analysis that varies the percentage of trials in which counterfactual reasoning is applied shows that counterfactual reasoning is key to explaining the human data. As discussed in the introduction, counterfactual reasoning is essentially learning by imagining the interactions between the organism and the outside world. Compared to learning by actually experiencing the world, it incurs almost no risks. Understandably, it might be a powerful tool that drives animals to safely explore novel options in response to unusual changes of the environment. Our research suggests that this is likely the case, and perhaps future theories and models of learning and decision making should always incorporate counterfactual reasoning.

Acknowledgments

This work is supported in part 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 opinions expressed hereon are strictly those of the authors.

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.

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

Coricelli, G., Critchley, H. D., Joffily, M., O'Doherty, J. P., Sirigu, A., & Dolan, R. J. (2005). Regret and its avoidance: a neuroimaging study of choice behavior. *Nature Neuroscience, 8*(9), 1255–1262.

Holroyd, C. B. & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109*(4), 679–709.

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

Meertens, R. & Lion, R. (2008). Measuring an individuals tendency to take risks: the risk propensity scale. *Journal of Applied Social Psychology, 38*(6), 1506–1520.

Nassar, M., Wilson, R., Heasly, B., & Gold, J. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *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.

Stephens, D. W. (1987). On economically tracking a variable environment. *Theoretical Population Biology, 32*(1), 15–25.