Information versus reward in a changing world

Daniel J. Navarro (daniel.navarro@adelaide.edu.au)
Ben R. Newell (ben.newell@unsw.edu.au)

Abstract

How do people solve the explore-exploit trade-off in a changing environment? In this paper we present experimental evidence in an "observe or bet" task, comparing human behavior in a changing environment to their behavior in an unchanging one. We present a Bayesian analysis of the observe or bet task and show that human judgments are consistent with that analysis. However, we find that people's behavior is most consistent with a Bayesian model that assumes a rate of change that is higher than the true rate in the task. We argue that this tendency is the result of asymmetric consequences: assuming that the world changes more often than it really does is not very costly, whereas assuming a too-low rate of change can carry much more severe consequences.

Keywords: decision making; explore-exploit dilemmas; learning; change detection

Introduction

"When one thinks about designing intelligent agents, it quickly becomes obvious that the task environment in which the agent will operate is a primary determiner of the appropriate design" – Jordan and Russell (1999, pp. lxxv)

"Outside of gambling casinos and psychology laboratories, there are few – if any – circumstances where one can safely assume conditional independence of a succession of events" – Ayton and Fischer (2004, pp. 1369)

Consider the problem facing someone who wants to make her first investments on the stock market. Initially she knows very little about what stocks to buy, and so must spend time learning about the market before making a purchase. Once the investments are made, external constraints (e.g., work, family) mean that she cannot afford to spend a lot of time monitoring investments: once the purchase is made, she is more or less required to ignore low level details of how the investments are performing most of the time. Occasionally, when she has time to re-evaluate her portfolio, she might look at the market in more detail to determine whether to change her investments. Given this, how much time should she spend researching her initial investment? How often should she revisit her portfolio to consider making changes? When she does so, how much time should she spend revisiting her original decisions?

The stock market scenario is an example of an *explore-exploit* problem. The actor is operating in an environment that can generate rewards (money) from different options (purchases), but is initially uncertain about which options are good and which are bad. To maximize rewards, some proportion of the actor's time must be spent obtaining information about the structure of the environment ("exploration"), and some proportion on using this knowledge to extract rewards ("exploitation").

Explore-exploit problems have been studied a lot in psychology (see Cohen, McClure, & Yu, 2007). In this paper we focus on the "observe or bet" task (Tversky & Edwards, 1966; Rakow, Newell, & Zougkou, 2010). Like the better known bandit problem (Robbins, 1952; Steyvers, Lee, & Wagenmakers, 2009), the observe or bet task provides an elegant experimental framework within which to study explore-exploit dilemmas, and can be viewed as a highly simplified version of the stock market scenario. The decision maker has a number of options available (stocks), each of which may yield rewards (or losses). On each trial, she may choose to observe the state of the world (i.e. do research), in which case she gets to see what rewards each option provided, but receives no reward nor suffers any losses. Alternatively she may pick one of the options and receive the rewards/losses associated with that option at the end of the task. However, when she does she receives no information at that time: the outcomes are hidden from her. Although a little unrealistic, the design is qualitatively in keeping with the constraints posed by the real world problem and the design of the task creates a clean separation between information gathering (observe) and reward taking (bet).

The focus of the paper is on how people make the required trade-off in a changing environment. Much of the literature focuses on problems in which the underlying structure of the learning problem is static (Shanks, Tunney, & McCarthy, 2002), presumably because the inference problem is simpler. Yet, as discussed by Ayton and Fischer (2004), the actual environments in which people have to operate are not static: things change. Strategies that are optimal in an unchanging world may be highly maladaptive in a dynamic world. If people are accustomed to making choices in a dynamic environment and use strategies that are appropriate to such an environment, studying their behavior using static tasks may be highly misleading. In addition to its relevance to explore-exploit problems (Cohen et al., 2007), this idea has been discussed in connection with classic decision making biases (Ayton & Fischer, 2004), sequential effects, (Yu & Cohen, 2008) and categorization (Navarro, Perfors, & Vong, 2013). In an explore-exploit context, people show sensitivity to changing reward probabilities in bandit tasks (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006).

Rational choice in the observe or bet task

Consider the simplest version of an observe or bet task. There are only two outcomes (e.g., "pick blue" or "pick red"), and on every trial one option provides a reward and the other provides a loss. The learner's goal is to use observations to determine which option is more likely to be the reward-generating one, and bet on that option. If the reward probabilities do

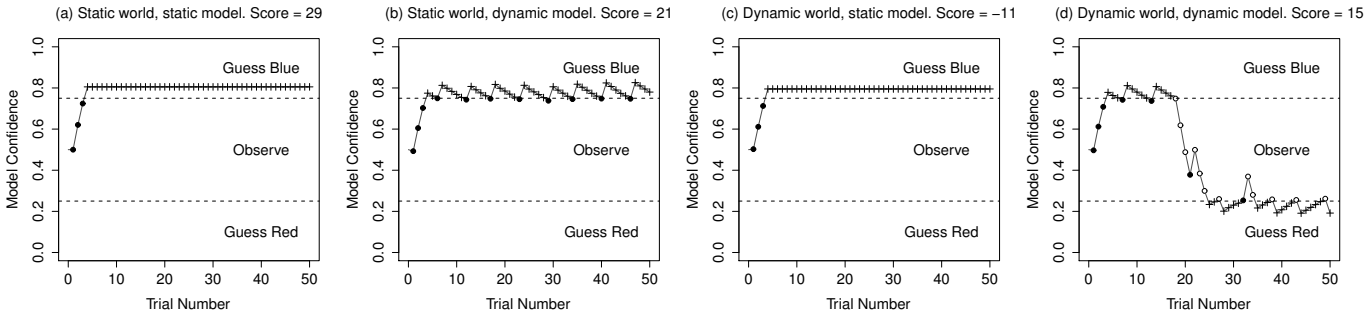

Figure 1: Model performance in a static world (panels a and b) and a dynamic one (panels c and d). The actual outcomes used correspond to the static and dynamic conditions in game 2, version 3 of the experiment (see Figure 2). In all cases the model requires a confidence of 75% in order to make a bet. Model performance is shown for an assumed change rate of 0% (panels a and c) and 5% (panels b and d). See main text for details.

not change over time, there is a well-known optimal strategy to solve the task, and previous studies have shown that people rarely follow it (Tversky & Edwards, 1966; Rakow et al., 2010). However, the analysis is not applicable to situations where the probabilities of different outcomes can change, or where the learner is uncertain about whether the environment is truly static. In this section we outline a new analysis of the observe or bet task that is more appropriate to a changing world. The analysis retains the core features of the original analysis, but uses the "dynamic belief" framework (Yu & Cohen, 2008) to allow the learner to respond appropriately to changes in the world.

The static task

The rational analysis presented by Tversky and Edwards (1966) is based on the sequential probability ratio test (SPRT: Wald, 1947), and is closely related to standard psychological models of choice reaction time experiments (e.g., Ratcliff, 1978). The SPRT describes the behavior of a rational learner who needs to choose between two hypotheses about data that arrive over time, using as few observations as possible. In the binary observe or bet task, the key variable to be inferred is θ, the probability that (say) "blue" will be the reward giving option. Let x_i denote the observations that the learner makes on trial i, where $x_i = 1$ means that the learner has observed a blue light, $x_i = -1$ means that the learner has observed a red light, and $x_i = 0$ means that the learner has not observed anything on this trial. The complete set of data available to the learner on trial t corresponds to the vector $\boldsymbol{x}_t = (x_1, x_2, \ldots, x_t)$. Using Bayes' rule we see that the posterior distribution over θ on trial t is given by

$$P(\theta|\boldsymbol{x}_t) \propto P(\theta) \prod_{i=1}^{t} P(x_i|\theta) \tag{1}$$

where $P(\theta)$ denotes the learner's prior over the bias[1] and

$$P(x_i|\theta) = \begin{cases} \theta & \text{if } x_i = 1 \\ 1 & \text{if } x_i = 0 \\ 1-\theta & \text{if } x_i = -1 \end{cases} \tag{2}$$

[1] We fixed this to be a Beta(5,5) distribution.

Of course, the primary goal in the task is not to infer the specific value of θ, but to determine which option to bet on. That is, the learner needs to infer whether $\theta > .5$ or $\theta < .5$. The relevant probability is

$$P(\text{choose "blue"}) = P(\theta > .5|\boldsymbol{x}_t) = \int_{.5}^{1} P(\theta|\boldsymbol{x}_t)\, d\theta \tag{3}$$

The SPRT strategy is essentially equivalent to this Bayesian analysis, in which the decision-maker specifies an allowable tolerance ε for making incorrect decisions.[2] Initially, the learner does not know which option is best, and so it starts by making observations, and updating his or her beliefs about the relative goodness of different options via Bayes' rule. Once the posterior probability of one of the two choices exceeds the required evidentiary standard (i.e., posterior probability exceeds $1-\varepsilon$), it then chooses that option on the next trial and for all future trials. It never makes new observations because it has absolute faith that nothing can change: if guessing "blue" was the right thing to do on trial 5, it will remain the right thing to do for all subsequent trials.

Adapting to change

What happens to this strategy if the world can change? As before, the model has a fixed tolerance for errors ε. It will choose a particular option on trial t only if the posterior probability that it is the correct choice exceeds $1-\varepsilon$, and will observe if no option reaches this threshold. When the world can change, the correct choice on (say) trial 5 may not be the correct choice on trial 50, and so the confidence in a chosen option will decrease over time if the model does not continue to receive confirming evidence.

The difficulty for the learner is that he or she can never be certain if a change has occurred. In the dynamic observe or bet task, just as in real life, changes can occur without warning: it is only by continuing to make observations and

[2] The SPRT approach remains agnostic about how ε should be set. It is possible to be more precise about the optimal strategy by treating the task as a partially observable Markov decision policy, an approach that has been used in related problems (e.g. Frazier & Yu, 2007; Zhang & Yu, 2013). To keep things simple, we forbear from introducing this additional complexity in this initial investigation.

comparing those observations to your expectations that you can infer that a change has happened. The observe or bet task is particularly difficult in this respect, because the learner is required to explicitly decide to give up a potential reward every time he or she wants to "check" that his or her betting strategy is still effective.

To formalize this intuition, we follow previous work (Yu & Cohen, 2008; Brown & Steyvers, 2009) and use a simple change model: the model assumes that on every trial there is some probability α that the world changes in an entirely arbitrary way. To incorporate this idea into the model the key observation is that in a static world, "today's priors are yesterday's posteriors". That is, the Bayesian posterior at time t in Equation 1 could equally have been described in terms of the posterior at time $t-1$, like so:

$$P(\theta|\boldsymbol{x}_t) \propto P(x_t|\theta)P(\theta|\boldsymbol{x}_{t-1}) \tag{4}$$

However, when things can change, this relationship no longer holds. If we let θ_t denote the probability that blue is the right choice on trial t, then the corresponding equation becomes

$$P(\theta_t|\boldsymbol{x}_t) \propto P(x_t|\theta_t) \int_0^1 P(\theta_t|\theta_{t-1})P(\theta_{t-1}|\boldsymbol{x}_{t-1})\, d\theta_{t-1} \tag{5}$$

where the term $P(\theta_t|\theta_{t-1})$ describes the dynamics of the world: that is, it describes the conditional distribution over the true probability at time t (i.e., θ_t) given that the true probability on the previous trial was θ_{t-1}. Specifically, with probability α there is no change and θ_t is thus identical to θ_{t-1}, but with probability α it changes unpredictably, and so the learner must assume that θ_t is a random sample from the prior distribution $P(\theta)$.

The resulting model turns out to be very tractable when particle filtering is used to compute posterior distributions over θ at every trial (Doucet, De Freitas, & Gordon, 2001; see Sanborn, Griffiths, & Navarro, 2010 for psychological applications). From these posterior distributions it is straightforward to infer when the model observes and when it bets.[3]

Illustrating model behavior

The behavior of the model is illustrated in Figure 1. Each plot shows the model's confidence (i.e., posterior probability that $\theta > .5$) on every trial. The markers indicate whether the model observed a blue light (black circle), or a red light (white circle), or whether the model chose to bet on that trial (crosses). If the model assumes that θ does not change over time (i.e. $\alpha = 0$, panels a and c), the model continues to make observations until some required degree of confidence is reached (dashed lines), at which point it begins to bet and continues to do so until the end of the game. When the model believes that θ can change over time ($\alpha = .05$, panels b and d), the behavior is quite different. As before, it observes for a time at the beginning of the task, at which point it begins to bet. However, because the model receives no information from a bet, the probability that an unobserved change has occurred rises, causing its confidence to decrease. Eventually, the confidence falls below threshold and it makes another observation. If the world has changed the model may detect it and change its betting strategy: this happens in panel d. The figure shows how well the model performs in different conditions if a point is earned for every correct guess, and one is lost for every incorrect guess. The dynamic model performs reasonably well if the world is static (21 points) or if it changes (15 points). The static model performs well when the world is static (29 points), but performs very poorly when the world can change (-11 points).

[3] R code implementing the model using particle filtering is available at `https://bitbucket.org/dannavarro/observe-or-bet`

Experiment

Previous research suggests that human behavior in static observe or bet tasks does not mirror the static version of the model, but has some similarity to the dynamic version (see Tversky & Edwards, 1966; Rakow et al., 2010): even when the world is static, people have a tendency to keep checking, just to make sure that the strategy they are following is still correct. The analysis presented above suggests that there may be a good reason for this: the consequences of incorrectly assuming the world is static appear to be more severe than those for incorrectly assuming that it is not. In this experiment we explore this idea, and use the model to infer how changeable people expect the task to be.

Method

Participants A total of 108 US-based workers (46 female) on Amazon Mechanical Turk participated in the experiment, randomly assigned to conditions. The task took about 10 minutes to complete and workers were paid US$0.60 for their time. Mean reported age was 34.9 (std dev = 11.7).

Materials & Procedure The task took the form of a guessing game, in which participants were told about "blox machines", which are devices that have two lights (blue and red). On every trial, one of the two lights would turn on, and the task was to predict which one it would be. The cover story explained that every blox machine has a bias to prefer one light or the other, but that their behavior was otherwise arbitrary. In the dynamic condition, participants were also told that sometimes the bias on a blox machine could randomly change, and that although such changes were rare they should expect to encounter a few of them during the experiment. The bias was always of the same magnitude (a 70:30 split), but participants were not informed of this.

The structure of the observe or bet task was then explained: on any given trial they could either choose to see what color light turned on (but receive no points), or they could guess a color. If they guessed, they would not be shown what actually happened. Nevertheless, they were told that they would receive one point for every correct guess, and lose one point for every incorrect guess.

Each participant played 5 observe or bet games (each with

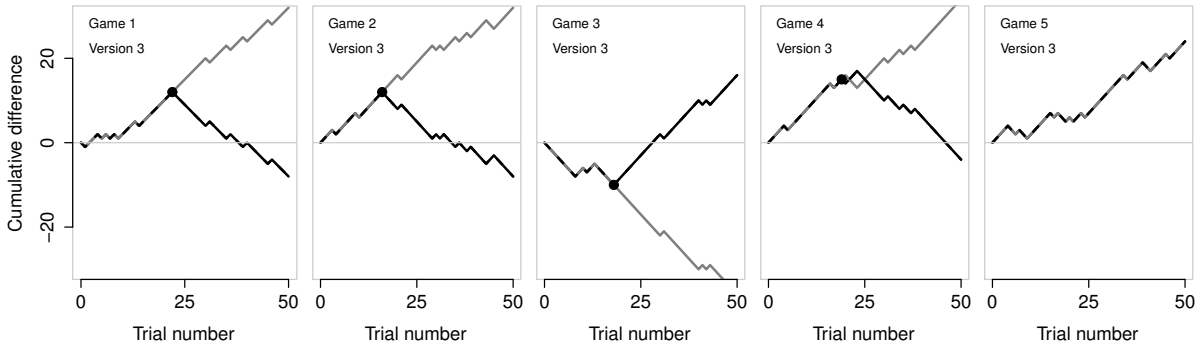

Figure 2: Outcome sequences used in one version of the experiment (version 3 only). Each sequence is shown as a random walk: every time the machine turns on the blue light (whether observed by the participant or not), the walk takes a step up, and every time it turns red the walk takes a step down. The sequences for the static condition are shown as the grey lines, and always have a consistent bias in one direction or the other. The dynamic sequences are initially identical to the corresponding static sequences. After a change occurs (circles) the dynamic sequences have the opposite outcomes to the corresponding static ones (black lines). Note that in game 5, the dynamic sequence did not have a change point, so the two conditions used identical data in this case. This also happened in version 1 game 1 and version 2 game 3.

a new blox machine) consisting of 50 trials, and at the end of each game received detailed feedback that showed what they did on every trial and what the machine did on every trial. It showed which trials they won points for and which they lost points for, as well as giving them a final score.

There were three versions of each condition, each with a fixed set of outcomes (see Figure 2). In all three versions, the dynamic condition involved a change in 4 of the 5 games, but in one game the dynamic and static outcomes were identical. By doing so, it was tested whether expectations about change affect people's behavior independently of whether any actual change occurs.

Results

The main variable of interest is how often people choose to observe rather than bet, and critically, when they choose to do so.[4] With this in mind, Figure 3a plots the proportion of trials on which participants chose to observe rather than bet, shown as a function of trial block, condition, and game number. In both the static and dynamic conditions there is a tendency, across all five games, for people to observe the most at the beginning of the game, and least towards the end. As predicted, this effect is more pronounced for the static condition than the dynamic condition.

To quantify this effect, it is useful to analyze the data from the first 10 trials (block 1) of each game separately from the other trials. For block 1, we fit a linear mixed effects model that included fixed effects of condition and game. For the remaining data, we fit a separate mixed model with fixed effects of condition, game and block. In both cases individual differences were captured with a random intercept term. During block 1, participants in the stationary condition were significantly more likely to observe than the participants in the dynamic condition (analysis of deviance: $\chi^2(1) = 5.00, p = .025$), whereas in blocks 2 to 5 they were less likely to do so ($\chi^2(1) = 9.24, p = .002$). The amount of observation in block 1 did not change as a function of game ($\chi^2(1) = .005, p = .94$), but for the later blocks people were less likely to observe in the later games ($\chi^2(1) = 52.3, p < .001$). It is obvious from inspection of the plots that people were more likely to observe during block 1 than later blocks: additionally, the statistical analysis confirmed that the smaller trend towards less observation from blocks 2 to 5 is also significant ($\chi^2(1) = 75.5, p < .001$).

One possibility to consider is that, although the pattern of responses in Figure 3a appear quite different, it might be that people are following the same strategy in both conditions. Any differences in the data might be caused by incidental differences in the sequences, and not due to any inferences people are making about the changing nature of the environment. Our design lets us test this directly, since there are three cases when the sequences are identical for the two conditions: game 1 in version 1, game 3 in version 2, and game 5 in version 3 (see Figure 2).

The data for those cases are plotted in Figure 4. There are no differences in responses on game 1 (version 1), implying that the cover story manipulation alone had no effect on people's behavior. The differences in behavior on game 1 in Figure 3a are driven by people adapting to the changes that actually do occur on game 1 for versions 2 and 3. In contrast, by the time people play games 3 and 5, there are genuine differences in performance even when the actual sequences are identical. Previous experience with changing (or static) environments has altered the approach that people take.

Modelling the observe or bet data

Model fitting was done by grid search over evidence thresholds and change rates, minimizing the sum squared error across the five blocks. The model was fit separately to each subject and each game. The average model behavior is plotted in Figure 3b. Across all blocks, games and conditions the model predictions correlate with human performance at

[4]The betting strategies people followed are also interesting: people sometimes probability match rather than select the best option given their observations. However, space constraints prevent us from discussing this aspect of the data set.

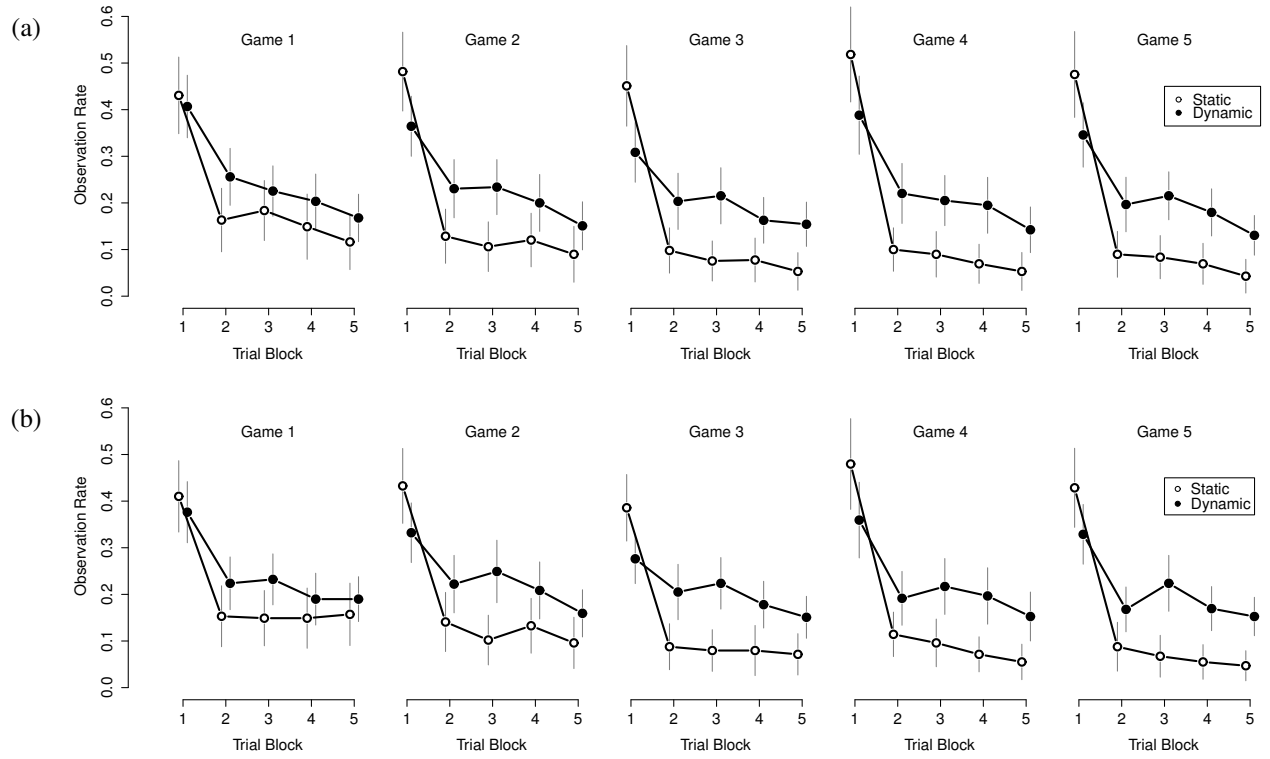

Figure 3: (a) Proportion of trials on which participants chose to observe rather than bet, shown as a function of trial block, condition, and game number. Each block consists of 10 trials. Error bars plot 95% confidence intervals. (b) Model behavior when fit to the human data.

$r = .97$ ($p < .001$). The correlation between model predictions and human data was significant in 97 of 108 cases, with an average correlation of $r = .83$.[5]

The parameter estimates are shown in Figure 5, and reveal two interesting facts. Firstly, the best fitting change rate parameters were higher in the dynamic condition than the static condition, as one would expect. The model fits imply that people in the dynamic condition were behaving (on average) as if they expected changes to occur on 7.5% of the trials, whereas in the static condition they behaved as if changes were expected 3.5% of the time. In both cases, the true rate of change was substantially lower than these estimates suggest. In the static condition, the true change rate was obviously 0%. In the dynamic condition, there were 4 change trials among the 250 trials each participant played, so the true change rate was only 1.6%. Both are much lower than the implied change rates that emerge from the model fitting.

Secondly, the evidence thresholds were higher in the stationary condition than in the dynamic condition: the model fits imply that in the stationary condition, an option needed to have a 71% chance of being correct in order for people to be willing to bet, whereas in the dynamic condition this falls to 63%. This may not be arbitrary: if the world can change, it may make sense to act quickly (with less evidence) to take rewards before it *does* change.

[5]Note that the model uses 10 parameters to fit the 25 blocks of data provided by each person, so the very good fits are only moderately strong evidence. We used this non-parsimonious version of the model because we wanted separate parameter estimates for each game (see Figure 5).

Discussion

Why study human behavior in changing environments? As a discipline we find it difficult enough to provide clear theoretical accounts of how people behave on static problems. Making the task more complex and dynamic might appear to be a recipe for disaster. The problem, as we see it, is that human cognition is adapted (either via evolution or prior learning) to operate in changing and responsive worlds. This matters: as noted by (Jordan & Russell, 1999), the structure of the operating environment imposes strong constraints on how an intelligent agent can be built. This basic idea underpins both the "fast and frugal heuristics" literature (e.g. Simon, 1956; Gigerenzer & Brighton, 2009) as well as the "rational analysis" approach to building cognitive models (e.g. Anderson, 1990). Viewed from this perspective, studying human cognition in an unchanging world is like studying the aerodynamic properties of an octopus: technically possible, because water and air are both fluids, but frustrating and misleading because it ignores what the organism is designed to achieve.

The observe or bet task is an instructive case. As illustrated by Figure 1, the "rational" strategy discussed by Tversky and Edwards (1966) turns out to be limited and very fragile. A learner who incorrectly assumes that the world is changeable and follows the strategy that is optimal on that basis will per-

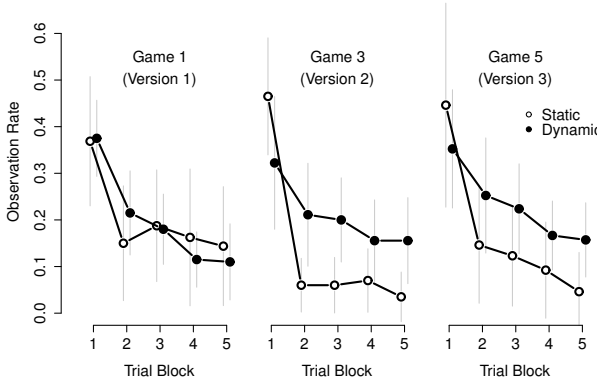

Figure 4: Proportion of trials observed broken down by trial block, for the three test cases. Sequence 1 in version 1 happened to be identical (and unchanging) in both conditions, as was sequence 3 in version 2 and sequence 5 in version 3. There appears to be no "pure" cover story effect, insofar as the responses on game 1 are the same in both conditions. In later games, however, an effect is observed that is qualitatively identical to the one shown in Figure 3. Given that the sequences are identical, this difference implies that the observed effect is a genuine strategy shift between conditions, and not merely the application of a single strategy that produces different results when given different data.

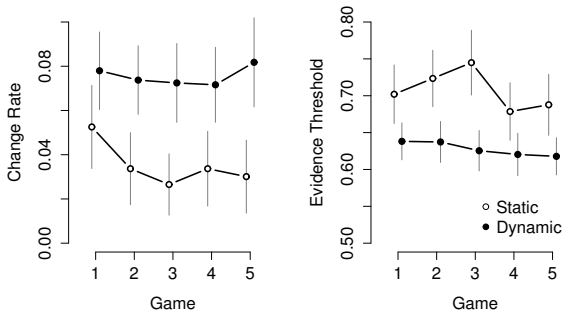

Figure 5: Estimated model parameters by condition and game. In the dynamic condition the subjective change rates are higher, and the evidence thresholds are lower.

form slightly worse in psychology experiments, and perhaps runs the risk of being called "irrational" by the researchers. On the other hand, a learner who incorrectly assumes that the world is static can perform catastrophically badly when the world changes.

This asymmetry has consequences for our data. When we fit the model to human data, we found that people act as if they expected a lot more changes than actually occurred. In the static condition (true change rate 0%) people expected a change rate of 3.5%. In the dynamic condition (true change rate 1.6%) people appeared to expect a change rate of 7.5%. Viewed narrowly, we might conclude that people are miscalibrated in their assessment of how to solve explore-exploit problems in changing environments. A more charitable interpretation would be to note that people can never know with certainty what the rate of change might be (even if we were to tell them: experimenters lie), and that one error is worse than the other. If underestimating the rate of change is a vastly more dangerous mistake than overestimating it, then an intelligent decision maker should err on the side of caution and act as if the rate of change is much higher than it actually is.

Acknowledgments

DJN received salary supported from ARC grant FT110100431 and BRN from ARC grant FT110100151. Research costs were funded through ARC grant DP110104949. We thank Tim Rakow, Amy Perfors and Nancy Briggs for helpful comments.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & Cognition*, *32*(8), 1369–1378.

Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, *58*(1), 49–67.

Cohen, J. D., McClure, S. M., & 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: Biological Sciences*, *362*(1481), 933–942.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

Doucet, A., De Freitas, N., & Gordon, N. (2001). *Sequential Monte Carlo methods in practice* (Vol. 1). Springer New York.

Frazier, P., & Yu, A. J. (2007). Sequential hypothesis testing under stochastic deadlines. In *Advances in neural information processing systems* (p. 465–472).

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143.

Jordan, M. I., & Russell, S. (1999). Computational intelligence. In R. A. Wilson & F. C. Keil (Eds.), *The MIT encyclopedia of the cognitive sciences* (pp. 1xxiii–xc). Cambridge, MA: MIT Press.

Navarro, D. J., Perfors, A., & Vong, W. K. (2013). Learning time-varying categories. *Memory & Cognition*, *41*, 1–11.

Rakow, T., Newell, B. R., & Zougkou, K. (2010). The role of working memory in information acquisition and decision making: Lessons from the binary prediction task. *The Quarterly Journal of Experimental Psychology*, *63*(7), 1335–1360.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of Australian Mathematical Society*, *55*, 527–535.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*(3), 233–250.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*(3), 168–179.

Tversky, A., & Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology*, *71*(5), 680–683.

Wald, A. (1947). *Sequential analysis*. Dover Publications.

Yu, A. J., & Cohen, J. D. (2008). Sequential effects: superstition or rational behavior? In *Advances in neural information processing systems* (p. 1873–1880).

Zhang, S., & Yu, A. J. (2013). Forgetful Bayes and myopic planning: Human learning and decision-making in a bandit setting. In *Advances in neural information processing systems* (p. 2607–2615).