Detecting Regime Shifts:
The Causes of Under- And Over-Reaction*

Cade Massey
George Wu

Fuqua School of Business
Duke University
Durham, NC 27708
cade.massey@duke.edu

Graduate School of Business
University of Chicago
Chicago, IL 60637
george.wu@gsb.uchicago.edu

June 4, 2004

ABSTRACT

Many decision makers operate in dynamic environments, in which markets, competitors, and technology change regularly. The ability to detect and respond to these regime shifts is critical for economic success. We conduct three experiments to test how effective individuals are at detecting such regime shifts. Specifically, we investigate when individuals are most likely to under-react to change and when they are most likely to over-react to it. We develop a *system-neglect hypothesis*: individuals react primarily to the signals they observe and secondarily to the environmental system that produced the signal. Three experiments, two involving probability estimation and one involving prediction, reveal a behavioral pattern consistent with our system-neglect hypothesis: under-reaction is most common in unstable environments with precise signals and over-reaction is most common in stable environments with noisy signals. We test this pattern formally in a statistical comparison of the Bayesian model with a parametric specification of the system-neglect model.

---

* We thank participants at several conferences and numerous workshops for their comments and suggestions. We particularly thank J.B. Heaton for introducing us to this area, and Bill Goldstein, Chip Heath, Rick Larrick, Jack Soll, Yuval Rottenstreich, three anonymous reviewers, and the Associate Editor for their valuable comments. We also thank Kaitlyn Hwang, David Maloney, John Morse, Amanda Snow, and Caroline Tse for help in collecting the data.

1. INTRODUCTION

Decision makers must often make decisions in unstable environments. An investor must elect whether to keep her investments in equity. A Central Bank must decide whether a weakening economy merits an interest rate cut. A conventional retailer must decide whether to embrace an Internet strategy. In each of these examples a decision maker receives a "signal" (*e.g*, a series of poor earnings announcements, an unexpectedly high unemployment figure, an onslaught of "e-tailers") and must judge whether that signal augurs a new regime (*e.g.*, the onset of a bear market, an economy headed toward recession, a shift to an online economy), or is just an extreme outcome from the incumbent regime.

The ability of decision makers to correctly identify the onset of a new regime can mean the difference between success and failure. Indeed, the difficulty of separating "signal" from "noise" has led to well-publicized instances of *over-reaction* (*i.e.*, believing a regime shift has occurred before it actually has), as well as *under-reaction* (*i.e.*, believing a regime shift has not occurred when in fact it has). Consider two examples. In the late-1980's, Xerox observed Japanese semiconductor manufacturers shifting resources to a new, X-ray-based manufacturing method. Xerox, believing this indicated a fundamental shift in semiconductor technology, reallocated significant company resources to this new approach. Years later it was clear that such a shift did not occur and, in fact, would not for the foreseeable future (Grove, 1999). Conversely, in the late-1970's, Schwinn was slow to respond to the advent of the mountain bike. Management believed that the surge in mountain bike popularity was just another "fad". Schwinn's reluctance to introduce a mountain bike cost the company its dominance in the American bicycle market and contributed to its eventual bankruptcy (Crown & Coleman, 1996).

For obvious reasons, these examples must be interpreted with caution. In order to understand the factors that impact the ability of individuals to detect regime shifts more precisely, we investigate under- and over-reaction experimentally. We conduct three studies and find that individuals exhibit systematic biases in their ability to detect regime shifts. Subjects in our studies observe signals, and based on these signals indicate whether there has been a shift from one regime to a second regime.

We employ an experimental paradigm in which the first regime is represented by an urn that contains more red than blue balls (a "red urn"), and the second regime is represented by an urn that contains more blue than red balls (a "blue urn"). At any time, the experiment may switch from drawing balls from the red urn to drawing balls from the blue urn. In this setup, three values are needed for determining the appropriate (Bayesian) reaction: (i) the *signal*: the sequence of red and blue balls observed; (ii) *signal diagnosticity*: the degree to which the red urn differs from the blue urn; and (iii) *transition probability*: the chance that the system will remain with the red urn, or switch to the blue urn. This simple experimental paradigm is intended to capture the essential features of many real-world situations. For an investor deciding whether to keep her investments in equity, the two states might be a "bull market" and "bear market". An earnings announcement is an informative but imprecise signal, with the informativeness of this signal varying across markets and over time. Finally, the market may vacillate between these two states, with the historical frequency of change captured by the transition probability.

We posit and find strong evidence for a *system-neglect hypothesis*: individuals pay inordinate attention to the signal, and neglect diagnosticity and transition probability, the aspects of the system that generated the signal. Moreover, we suggest that *system neglect* leads to a predictable pattern of under- and over-reaction: individuals are most prone to under-reaction in unstable environments with precise signals and to over-reaction in stable environments with noisy signals. Our three studies, two judgment tasks and a choice task, show that subjects do exhibit substantial system neglect and that such system neglect does indeed result in the posited pattern of under- and over-reaction.

The paper proceeds as follows. In Section 2, we review previous empirical research on the detection of regime shifts, as well as relevant research on judgment in stationary environments. We extend these ideas and develop the system-neglect hypothesis. We test this hypothesis in three studies. In Section 3, we present a judgment study in which subjects judge the posterior probability that the process has switched regimes (from the red urn to the blue urn). In Section 4, we present a second judgment study that rules out error as an explanation for our results. In Section 5, we present a choice study in which subjects predict the next observation. In all three studies we find strong

evidence of system neglect, a lack of sensitivity to the system characteristics, diagnosticity and stability, and the predicted pattern of over- and under-reaction. We provide formal support for this insensitivity by estimating a family of "quasi-Bayesian" models. We conclude in Section 6 with a discussion of the implications of the research and an outline of future work.

2. BACKGROUND

Previous Research

We briefly outline some of the major empirical findings in the study of regime-shift detection. The studies have varied considerably in methodology. In some studies, the sequences were presented sequentially (Robinson, 1964; Chinnis & Peterson, 1968, 1970; Barry & Pitz, 1979), while in other studies, a sequence was provided in its entirety (Theios *et al.*, 1971). Some subjects estimated when the process changed from one regime to another (Barry & Pitz, 1979; Theios *et al.*, 1971), whereas other subjects estimated the probability that the data were drawn from one regime or another (Chinnis & Peterson, 1968, 1970). Studies have also varied environmental parameters such as diagnosticity (Robinson, 1964; Chinnis & Peterson, 1968, 1970), the payoff structure (Barry & Pitz, 1979), and the rate at which subjects receive stimuli (Robinson, 1964).

These studies have shown that individuals generally respond to the possibility of change and do so in the right direction. Beyond that, though, results are mixed. Across these studies, some behavior is approximately optimal (Chinnis & Peterson, 1968), but most is not (Barry & Pitz, 1979; Theios *et al.*, 1971). Studies find both under-reaction (Barry & Pitz, 1979; Chinnis & Peterson, 1968) and over-reaction (Brown & Bane, 1975; Chinnis & Peterson, 1970; Estes, 1984). One of the few themes connecting this literature is the idea that individuals respond *partially* to changing environmental conditions. Chinnis & Peterson (1968) state: "(t)he subjects, while sensitive to the difference in diagnostic value of the data in the two conditions, were not adequately sensitive" (p. 625). For example, Barry & Pitz (1979) systematically vary payoffs and find individual behavior largely unchanged. This theme resonates with the system-neglect hypothesis we develop in the next section.

Rapoport *et al.* (1979) provide a more comprehensive treatment of change detection, formally

modeling the decision environment and evaluating optimal policies. They develop and test three descriptive models in which individuals act only when the signal provides evidence exceeding some "threshold". Critically, these models incorporate extreme versions of system neglect that do not reflect *any* sensitivity to environmental conditions. For example, one model recognizes a change when the signal exceeds some fixed probability threshold.[1]

We suggest that the threshold model is too extreme. Instead, we hypothesize that individuals respond to changes in environmental conditions, but insufficiently so. In other words, their behavior is dictated more by signals than by the system generating the signals. We call this the *system-neglect hypothesis*. In the next section, we show how system neglect leads to under-reaction in some conditions and over-reaction in others.

Research Hypothesis

Our system-neglect hypothesis draws on judgment and decision making research in static environments. Our specific point of departure is Griffin & Tversky (1992), who reconciled two well-established, but seemingly contradictory, findings: conservatism and representativeness. Conservatism posits that individuals update their beliefs too slowly in the face of new evidence (Edwards, 1968; Grether, 1980). Representativeness, on the other hand, suggests that individuals extrapolate too readily from small samples, leading to belief revisions that are too dramatic (Kahneman & Tversky, 1973). Griffin & Tversky showed that sample size can account for this apparent contradiction. Most conservatism studies involve large samples, while most representativeness studies involve smaller samples. Misunderstanding the impact of sample size on the posterior probability leads subjects to make conservative revisions with large samples and radical revisions with small samples.

More generally, Griffin & Tversky distinguished between the *strength* and *weight* of evidence. Informally, the strength of evidence is its magnitude, while the weight of the evidence is its reliability. Imagine you are asked which way a coin is biased, 70% heads or 70% tails. Here, the proportion of heads represents the strength of evidence, while the sample size represents the weight

[1] This modeling approach is similar to that taken by Barry & Pitz (1979), Brown & Bane (1975), Estes

of evidence. Thus, 4 heads out of 5 flips has high strength but low weight, while 32 heads out of 60 flips has low strength but high weight. Griffin & Tversky showed that individuals systematically overweight the strength of evidence or how well that evidence matches the hypothesis in question, and underweight the weight of evidence or the diagnosticity of the signal. As a result, 4 heads out of 5 is seen to be more compelling than 32 heads of out of 60, contrary to Bayes Rule (since the difference between the number of heads and tails is a sufficient statistic).

We apply this idea to the problem of regime-shift detection. In such problems, an individual receives signals that are generated from one of two regimes. As in static environments, a decision maker must consider the diagnosticity of the signal, *i.e.*, how strongly correlated the signal is with the regime that it favors. In a dynamic environment, a decision maker must also take into account the transition probability, *i.e.*, how likely it is to change from one regime to another. Although a judgment about the likelihood of change should depend on both diagnosticity and transition probability, we suggest that individuals will not incorporate these two system parameters optimally. Instead, our *system-neglect hypothesis* posits that individuals respond primarily to the signal and secondarily to the system that generated the signal. In the parlance of Griffin & Tversky (1992), the signals provide the strength of evidence and the system parameters provide the weight of evidence.

A major reason that the signal is overweighted is that it is more salient: the signal is usually in the foreground while the system parameters are in the background. In many situations, the system parameters are not known and perhaps unknowable. In this sense, the system-neglect hypothesis is akin to the correspondence bias or fundamental attribution error in social psychology (Jones & Harris, 1967): Individuals tend to overestimate the extent to which behavior is due to disposition or personality and underweight the extent to which behavior is caused by the underlying situation.

The system-neglect hypothesis yields an important prediction: overemphasizing the strength of evidence at the expense of its weight means that individuals are most likely to *under*-react to an indication of change when the weight of that evidence is high and *over*-react when the weight is low. The weight of evidence is highest when diagnosticity is high and the system is unstable (*i.e.*, high transition probability), and lowest when diagnosticity is low and the system is stable (*i.e.*, low

---

(1984), and Robinson (1964).

transition probability). Thus, individuals should be most responsive to indications of change in precise (high diagnosticity) and unstable (high transition probability) environments and least responsive in noisy (low diagnosticity) and stable (low transition probability) environments. If individuals behave similarly across systems, we should see a pattern with relatively more *under*-reaction in precise/unstable environments and relatively more *over*-reaction in noisy/stable environments. Griffin & Tversky (1992) apply a similar logic and show that a strength and weight account predicts overconfidence when evidence has high strength and low weight (*e.g.*, low diagnosticity or small sample size), and underconfidence when evidence has low strength and high weight (*e.g.*, high diagnosticity or large sample size).

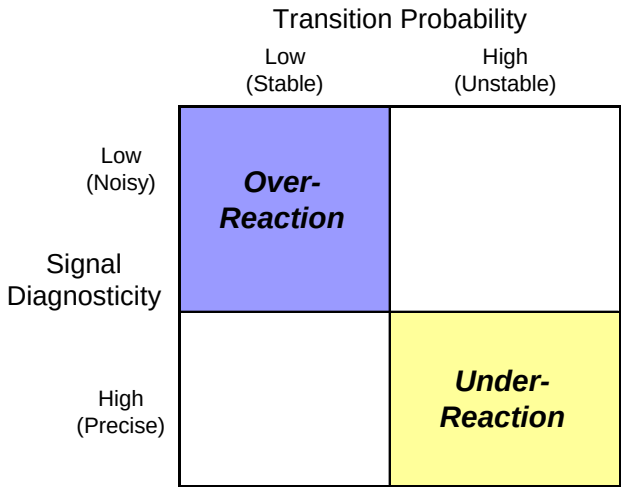


Figure 1: The system-neglect hypothesis: Predicted pattern of over- and under-reaction as function of system parameters

Figure 1 depicts these predictions succinctly. It is important that our prediction is a relative one: there should be relatively more over-reaction in the northwest cell, and relatively more under-reaction in the southeast cell. The hypothesis is silent about absolute levels. Indeed, the system-neglect hypothesis is consistent with over-reaction everywhere, under-reaction everywhere, or a mixed pattern, provided that the gradient slopes in the predicted direction. Note also that the predictions of Figure 1 apply to indications-of-change. Reactions to indications-of-*no*-change (or "business as usual") should not follow this pattern. Rather, under-reaction to indications-of-no-change should be strongest in unstable systems with imprecise signals, since the base-rate of change is highest and the signal least informative.

We test the system-neglect hypothesis in three studies by manipulating diagnosticity and transition probability and varying the task participants face.

3. STUDY 1: JUDGMENT TASK

Our experimental setting allows us to compare individual performance against a normative standard and thus test our behavioral predictions. The decision makers in our experimental paradigm have all the information necessary to calculate Bayesian responses and hence provide optimal judgments. This allows us to focus our analysis very specifically on the manner in which individuals revise probability judgments, and how these judgments deviate from Bayesian updating.

Statistical Process

We begin by describing the statistical process used in our experiment. The statistical process is depicted in Figure 2. Subjects observe signals generated by one of two regimes. The task is to detect if and when the system shifts from one regime to the other. The two possible regimes are red and blue. Let R_t (B_t) indicate that the process is in the red (blue) regime in period *t*. Each of the two regimes produces two possible signals: a red ball or a blue ball. While both regimes can produce either color, the red regime favors red balls (producing them with probability $p_R > 0.5$) and the blue regime favors blue balls (producing red balls with probability $p_B < 0.5$). In our studies, these probabilities are always symmetric, *i.e.*, $p_R = 1 - p_B$. Thus, p_R / p_B is a measure of the *diagnosticity* (*d*) of the signal, with a higher value indicating a more diagnostic signal.

The transition between the two regimes is defined by three characteristics. First, in period 0, before the first signal is drawn, the process is in the red regime ($\Pr(R_0) = 1$). Second, before each signal is produced, including the first one, the process might switch to the blue regime. The probability of such a switch is the *transition probability*, $q = \Pr(B_{t+1} \mid R_t)$. (This is called a "switch probability" in our experimental stimuli.) Third, the blue regime is an absorbing state, *i.e.*, if the process switches to the blue regime it remains there until the end of the trial ($\Pr(B_{t+1} \mid B_t) = 1$).

Again recall the investor example from our introduction. Diagnosticity captures the informativeness of a signal such as an earnings announcement, while the transition probability

captures the stability of a bull or bear market, *e.g.*, the likelihood that a bull market turns into a bear market. Although the absorbing state is a simplifying assumption in our model, for many decision horizons there will be at most one change. For example, most equity markets cycle between bull and bear markets at a sufficiently low frequency that for a given decision one can (and probably *should*) think of the new regime as an absorbing state. Of course, there are many other situations in which there is truly an absorbing state (*e.g.*, an equipment failure, an obsolete technology, etc.).

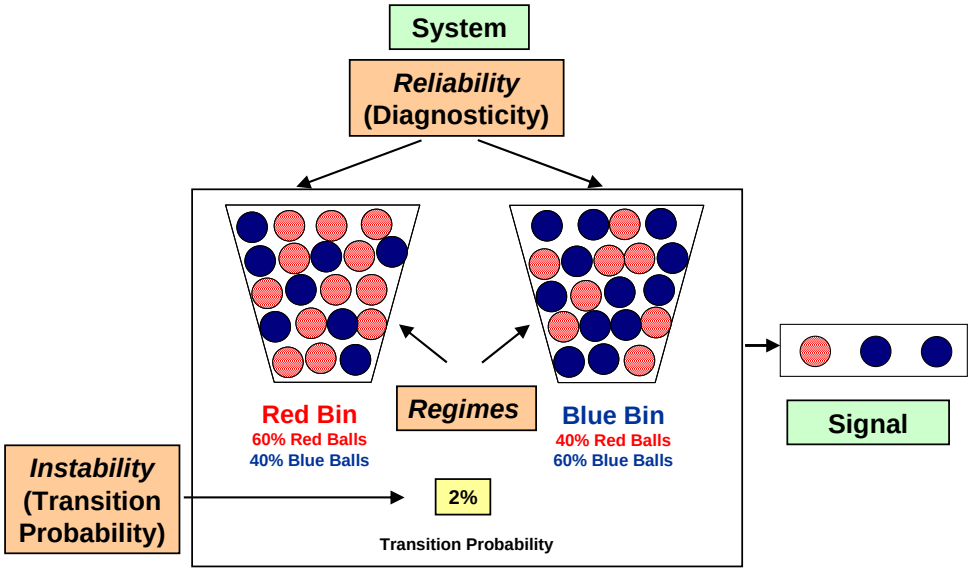


Figure 2: Statistical Process

Methodology

Study 1 was conducted on computer using a specially designed Visual Basic program. We recruited 40 University of Chicago subjects, advertising the task as a "probability estimation task" in order to yield subjects comfortable with probability. The median number of undergraduate and graduate mathematics and statistics classes taken by each subject was three.

The computer program began by introducing the statistical process used in the experiment, and illustrated the process using 4 demonstration trials and 2 practice trials. Detailed instructions and screen shots of the program are found in the Electronic Companion Paper (henceforth, ECP). Following the introduction, each subject completed 18 trials consisting of 10 signals (periods). Each trial was governed by a different set of parameters (see below). Subjects were first shown the parameters, p_R, p_B, and q, governing that trial. These parameters were displayed continuously

throughout each trial, and subjects were told that the parameters would change across trials. They were then shown a sequence of red or blue balls drawn randomly based on the set of parameters. After seeing each signal, subjects indicated the probability that the last ball was drawn from the blue regime (*i.e.*, the probability that the statistical process had *changed*). Subjects were not allowed to change a probability once it was entered.

We used 3 different diagnosticity levels and 4 different transition-probability levels, yielding 12 experimental conditions. Diagnosticity levels ($d = p_R / p_B$) were 1.5, 3, and 9, corresponding to (p_R, p_B) values of $(.6,.4)$, $(.75,.25)$, and $(.9,.1)$. The transition probabilities were .02, .05, .10, and .20. We chose the diagnosticity levels and transition probabilities to span a "reasonable" range of the parameter space. For example, 3 of 15 sequences show a regime shift for $q = .02$, while 12 of 15 sequences do so for $q = .20$, so moving the transition probability in either direction would yield a process that either always or never produced regime change during our series. We randomly generated 5 unique sequences for each condition, using a statistical process with the true parameter values, yielding 60 sequences in total (the actual sequences are found in ECP). Each of the subjects received 18 of the 60 sequences in a randomized order, and at least one sequence from each of the 12 conditions. Each of the 60 sequences was judged by a total of 12 subjects.

We paid subjects according to a quadratic scoring system that paid \$0.10 maximum (*e.g.*, if a subject indicated with certainty that the process was in the blue regime, and the process was in fact in the blue regime) and -\$0.10 minimum (*e.g.*, if a subject indicated with certainty that the process was in the blue regime, and the process was in fact in the red regime). Such a scheme is proper, and thus truth-revealing for risk-neutral subjects (Brier, 1950).

Subjects were given feedback at the end of each trial about if and when the process shifted from the red regime to the blue regime. They were also informed how much money they made or lost on that particular trial.

Normative Model

Our dependent measures are the probability judgments provided by subjects. We evaluate these probabilities by comparing them to a Bayesian standard. To do so, we derive the Bayesian

solution for our experimental framework (see Appendix). Let r_t denote the *t*-th signal, where $r_t = 1$ ($r_t = 0$) if a red (blue) ball is drawn in period *t*, and $H_t = (r_1,...,r_t)$ be the sequence of signals through period *t*. The Bayesian posterior odds of a change to the blue regime after observing history H_t can be expressed as

$$\frac{p_t^b}{1-p_t^b} = \frac{\Pr(B_t \mid H_t)}{\Pr(R_t \mid H_t)} = \left(\frac{1-(1-q)^t}{(1-q)^t}\right)\sum_{j=1}^{t}\frac{q(1-q)^{j-1}}{1-(1-q)^t} d^{\,t+1-j-\left(2\sum_{k=j}^{t} r_k\right)}, \tag{3.1}$$

where p_t^b denotes the Bayesian probability that the process has switched to the blue regime by *t*. The expression can be decomposed into several components. The first component, $\left(1-(1-q)^t\right)/(1-q)^t$, is a function only of the transition probability *q* and the number of signals *t* and provides a base rate for the likelihood of a change, *i.e.*, the likelihood of change in the absence of any data. The last component, $d^{\,t+1-j-\left(2\sum_{k=j}^{t} r_k\right)}$, is simply the diagnosticity raised to an exponent reflecting the difference between number of blue balls and red balls drawn over the part of the history from period *j* to period *t*. (This last component determines the posterior probability in a stationary environment [*cf.* Edwards, 1968].) The middle component provides mathematical weights accounting for the various paths through which the process can change from the red regime to the blue regime.

Experimental Results

We present both aggregate and individual-level results. Recall that our experiment required that each of our 40 subjects provide subjective probabilities that the process had switched to the blue regime for each of 10 signals in 18 trials. Consider first an overview of these judgments. Let p_t^e be the empirical judgment and $| p_t^e - p_t^b |$ be the absolute difference between the subject's judgment and the Bayesian probability for signal *t*. The mean absolute difference was .17 (median=.08, sd=.21). Subjects' payments were based on the difference between their subjective probability of a change to the blue regime and the actual regime (1 if blue, 0 if red). The mean of this absolute difference was .25 (median=.05, sd=.34), generating an average payment of $11.62 (median=$11.99, range of $6.60

to \$14.67). The average payment to a Bayesian agent would have been \$14.23 (median=\$14.22, range of \$12.72 to \$15.38).

Our primary interest, however, is belief *revision* – how probability judgments respond to new signals. Thus, we consider *changes* in probabilities rather than absolute levels and compare empirical changes in probability judgments with normative changes. Defining *empirical change* is straightforward: $\Delta p_t^e = p_t^e - p_{t-1}^e$. Constructing a proper normative measure of change requires a bit more care. In calculating the normative response to a signal *t*, we take a subject's previous probability judgment p_{t-1}^e as the "prior" rather than the previous Bayesian probability, p_{t-1}^b. If we use Bayes Rule with p_{t-1}^e as the "prior", we get

$$\frac{\overline{p}_t^b}{1-\overline{p}_t^b} = \frac{p_{t-1}^e}{1-p_{t-1}^e}\left(\frac{1}{1-q}\right)\left(\frac{p_R}{p_B}\right)^{1-2r_t} + \left(\frac{q}{1-q}\right)\left(\frac{p_R}{p_B}\right)^{1-2r_t}, \tag{3.2}$$

where $\overline{p}_t^b$ is the Bayesian response to signal r_t taking p_{t-1}^e as the prior.

This redefinition allows our "normative change" measure to reflect a subject's belief prior to receiving a signal. Thus, regardless of the accuracy of a subject's prior belief, our procedure automatically adjusts to that prior in order to evaluate the subject's belief *revision*. If, for example, a subject's previous judgment was low relative to the Bayesian probability, she should be "allowed" to make a more dramatic revision in light of the new signal (indeed, she is expected to). Conversely, if a subject's previous judgment was too high relative to Bayes Rule, there is less room for her to subsequently increase her judgment and therefore she should make a more conservative revision. Since the objective is to focus on the revision itself, we take as the normative change measure, $\Delta p_t^b = \overline{p}_t^b - p_{t-1}^e$, and compare this measure to the empirical change measure, $\Delta p_t^e = p_t^e - p_{t-1}^e$.

We restrict our attention to median judgments to minimize the role asymmetric error may play in generating system neglect (Erev, Wallsten, & Budescu, 1994). Erev *et al*. offer an error account in which responses are unbiased "true judgments" perturbed by error. This account requires that the median response be unbiased. In Study 2, we control for error more systematically. We also

report mean judgments in ECP.[2]

We first show that subjects are attentive to differences in signals. As expected, the median empirical change, Δp_t^e, is considerably different for blue signals than red signals, .084 versus .001. The median normative change measures are .158 and –.028 for blue and red, respectively. Thus, the median difference between the empirical change and normative change, $\Delta p_t^e - \Delta p_t^b$, our summary measure of over- and under-reaction, is slightly negative for blue signals (–.074), indicating a small overall tendency to under-react to indications of change.

To test for system neglect, we consider how our measure of under-reaction, $\Delta p_t^e - \Delta p_t^b$, varies across conditions. Recall that the predicted gradient shown in Figure 1 applies only to indications of change, thus we restrict our attention to blue signals. The system-neglect hypothesis implies that under-reaction is more common when diagnosticity and transition probability are high than when diagnosticity and transition probability are low. Indeed, the median normative change, Δp_t^b, is .231 in the high transition-probability/high diagnosticity condition, and .056 in the low transition-probability/low diagnosticity condition. In contrast, the median empirical change, Δp_t^e, across the two conditions is much less differentiated, .116 and .022, respectively.

Figure 3 depicts the difference between the empirical and normative change measures in each of our 12 experimental conditions. Note first that we observe under-reaction in 11 of the 12 conditions and over-reaction in 1 of the 12 conditions. By itself this pattern does not provide evidence for or against the system-neglect hypothesis. What matters is the gradient between the southeast cell (high diagnosticity and transition probability) and the northwest cell (low diagnosticity and transition probability). Consistent with system neglect, the greatest under-reaction occurs in the southeast-most cell, while the second least under-reaction occurs in the northwest-most cell. In the majority of the pairwise comparisons (37 of 48), the pattern of under-reaction is monotonic as

[2] We calculate the median change for each of the 600 observations (60 trials and 10 observations per trial) and then take the average across the appropriate category. For example, to calculate the median empirical change for a red ball, we take the median empirical change for the 329 (of 600) observations in which the signal is a red ball. We then take the average of these 329 medians. This method reflects the lumpiness with which subjects use the response scale. Taking the average over the medians provides us with a "smoother" measure.

transition probability and diagnosticity increase, as predicted by system neglect. The lack of perfect monotonicity partially reflects stimuli which are randomly generated and hence not necessarily "representative" of the underlying process.

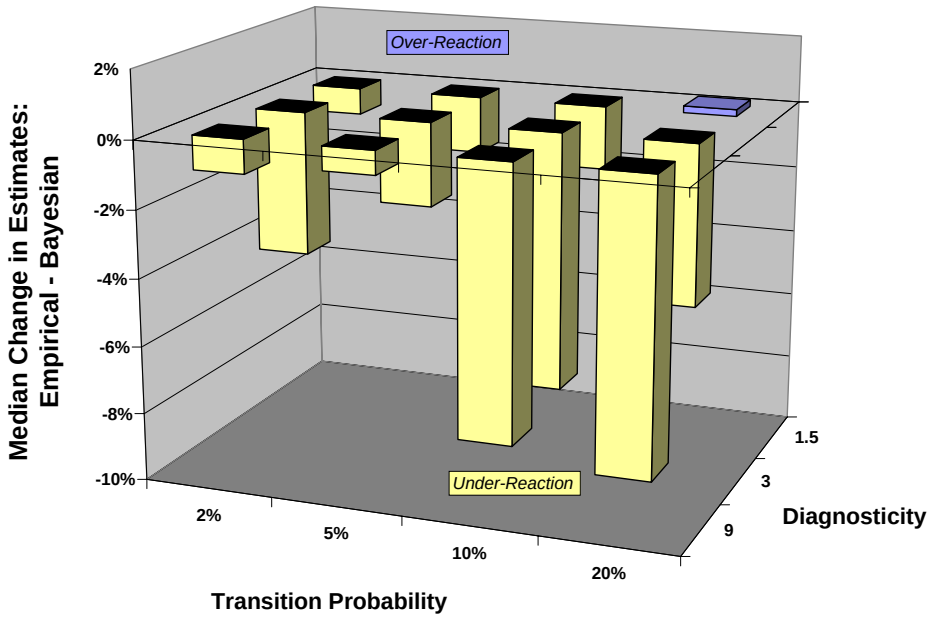


Figure 3: Over- and under-reaction, by condition, as measured by the median difference between change in empirical probability judgments and change in Bayesian probabilities, $\Delta p_t^e - \Delta p_t^b$ (Study 1, Judgment Task).

Estimation

The pattern in Figure 3 provides support for the system-neglect hypothesis. It is possible but unlikely that the hypothesized pattern reflects artifacts in the randomly generated sequences. Thus, we provide a formal test of the system-neglect hypothesis by generalizing (3.1). We call these generalizations *quasi-Bayesian models* in the spirit of Edwards (1968), who examined signal sensitivity in his study of conservatism (see also Chinnis & Peterson, 1968, 1970). The models are Bayesian in structure, but include additional parameters to accord more closely with the empirical observations.

In order to model sensitivity to the two critical dimensions, transition probability and diagnosticity, we add two parameters, α and β, to our Bayesian expression (3.1). Adding these parameters yields

$$\frac{p_t^e}{1-p_t^e} = \frac{\Pr(B_t \mid H_t)}{\Pr(R_t \mid H_i)} = \left(\frac{1-(1-\alpha q)^t}{(1-\alpha q)^t} \right) \sum_{j=1}^{t} \frac{q(1-q)^{j-1}}{1-(1-q)^t} d^{\beta\left[t+1-j-\left(2\sum_{k=j}^{t} r_k\right)\right]}, \tag{3.3}$$

where p_t^e is the subject's probability judgment that the process has switched to the blue regime by *t*. In (3.3), α captures sensitivity to transition probability, while β captures sensitivity to diagnosticity. This specification has several useful properties. The expression is Bayesian when $\alpha = 1$ and $\beta = 1$. As α shrinks, the transition probability plays an increasingly small role (as $\alpha \to 0$, $p_t^e \to 0$). As β shrinks, the signal has an increasingly small impact (as $\beta \to 0$, $p_t^e/(1-p_t^e) \to (1-(1-\alpha q)^t)/(1-\alpha q)^t$, *i.e.*, the "base rate" odds of change in the absence of information). We term the model in (3.3) the *power model*.

The formulation in (3.3) is well-suited for evaluating the degree of conservatism in belief revision (*cf.* Edwards, 1968). However, the power model is inadequate for evaluating system neglect, because it explicitly assumes that parameter values are constant across different systems. System neglect, on the other hand, suggests that these parameter values will vary systematically, reflecting insensitivity to system changes. To clarify the limitations of the power model, consider a more general formulation in which β depends on the diagnosticity condition *n*, $d_n^{\beta_n}$. For example, in the Bayesian model $\beta_n = 1$ for all *n*, implying that $d_n^{\beta_n}$ increases linearly as diagnosticity increases. In the power model, β may be less than 1 (as in conservatism) or greater than 1 but must be constant across all conditions. Thus, the power model is actually quite sensitive to changes in environmental conditions. Indeed, the responsiveness in the power model increases proportionally with the underlying parameter, with the proportion determined by the degree of conservatism. Note also that this model restricts behavior to be exclusively conservative (if $\beta_n < 1$) or exclusively "radical" (if $\beta_n > 1$), and thus permits *only* under-reaction *or* over-reaction, respectively.

The system-neglect hypothesis predicts that parameter values differ systematically in each condition. To illustrate, consider the extreme case of *complete* system neglect, in which individuals respond to signals identically regardless of the system producing the signals. For β, complete system

neglect requires that $d_1^{\beta_1} = d_2^{\beta_2} = d_3^{\beta_3}$. In our experimental design, $d_1 = 1.5$, $d_2 = 3.0$, and $d_3 = 9.0$. Thus, the equality required by *complete* system neglect implies that $\beta_1 = k$, $\beta_2 = .37k$, and $\beta_3 = .18k$, where k is any positive constant. Of course, this illustrates only the extreme version of system neglect. More generally, system neglect requires that for $q_m < q_n$ and $d_m < d_n$, parameter estimates be ordered so that $\alpha_m > \alpha_n$ and $\beta_m > \beta_n$.

To evaluate this prediction, we generalize (3.3) as follows. Let

$$\alpha = \alpha_1 Q_1 + \alpha_2 Q_2 + \alpha_3 Q_3 + \alpha_4 Q_4 \text{ and } \beta = \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3, \tag{3.4}$$

where Q_m is a dummy variable corresponding to transition probability q_m, and D_n is a dummy variable corresponding to diagnosticity condition d_n. This approach explicitly allows α_m and β_n to vary across environmental conditions, thus providing a formal test of the system-neglect hypothesis. The power model is a special case of this model in which $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$ and $\beta_1 = \beta_2 = \beta_3$, and the Bayesian model is a special case of the power model in which $\alpha_m = \beta_n = 1$ for all m, n.

Note that while the system-neglect hypothesis does not make any predictions about the overall level of conservatism in a particular environment, it does make explicit predictions, via the monotonic ordering of α_m and β_n, about the pattern of over- and under-reaction. The system-neglect model explicitly allows for mixed patterns of over-reaction ($\alpha_m, \beta_n > 1$ for some m, n) and under-reaction ($\alpha_m, \beta_n < 1$ for some m, n).

We fit p_t^e, using the model in (3.3) and manipulating the odds form. We use nonlinear regression with the usual assumption that errors are normally distributed around zero.[3] We estimate this model for each of the 600 observations using the median judgment for each observation. We also estimate the model for each of the 40 individual subjects.

We estimate both the "power model" and the "system-neglect model" (3.4). The estimates for the models are shown in Figure 4. Recall that system neglect predicts $\alpha_m > \alpha_n$ and $\beta_m > \beta_n$ for $m > n$. These orderings provide 9 pairwise comparisons as empirical tests of the system-neglect

[3] Residual analysis indicates this is a good assumption.

hypotheses, 6 for the α-coefficients and 3 for the β-coefficients. All 9 pairs of estimates are in the direction predicted by system neglect, and all are significantly different (all comparisons, $p < .01$).[4] Note also that parameter values are both greater than and less than one. This implies the coexistence of conservative and radical belief-revision, an important fact concealed by the power model.

There are three inferences to draw from Figure 4. First, comparing the *level* of the system neglect plot with the Bayesian plot reveals the overall level of conservatism: lower levels imply greater conservatism. Second, comparing the slope of the system-neglect model with the Bayesian model and complete-neglect models reveals the degree of system neglect: the steeper the slope, the more system neglect. The parameter estimates are close but distinguishable from complete neglect. Finally, comparing the system-neglect model with the power model reveals the benefit of using the system-neglect model: greater differences imply greater value. Figure 4 shows system neglect for both parameters.[5]

We also estimated the system-neglect model for individual subjects. Overall, this model fits the individual-level data reasonably well, with most individual-level parameter estimates ordered as predicted. Recall that the system-neglect hypothesis suggests that as system parameters increase, model estimates will decrease, suggesting 9 pairwise comparisons (6 for the α-coefficients and 3 for the β-coefficients). Across all subjects, 79% of the α-coefficients and 77% of the β-coefficients are ordered in the predicted direction.

Finally, we evaluated performance at different points in time to consider the effect of experience. Recall that subjects experience 18 trials over the course of the experiment (in addition to two practice trials) and thus might learn to avoid system neglect. We divide the session into four "quarters" comprising of 4, 5, 5, and 4 trials, estimating system-neglect models for each quarter separately. Because behavior in the final 3 quarters appears similar and learning appears limited to the first quarter, we aggregate those 3 quarters and compare them against the first quarter. While we

[4] The fits of the models are as follows: power model ($R^2 = .86$), system-neglect model ($R^2 = .95$), and Bayesian model ($R^2 = .84$). Since the purpose of the estimation is to test our psychological hypothesis, we do not emphasize the goodness of fit.

find significant system-neglect in both periods, it is more pronounced in the first quarter. Specifically, we observe much greater sensitivity to low transition-probabilities in the first quarter than in the subsequent quarters. Nevertheless, significant system-neglect persists throughout the experiment. Parameter values for each quarter are reported in ECP.

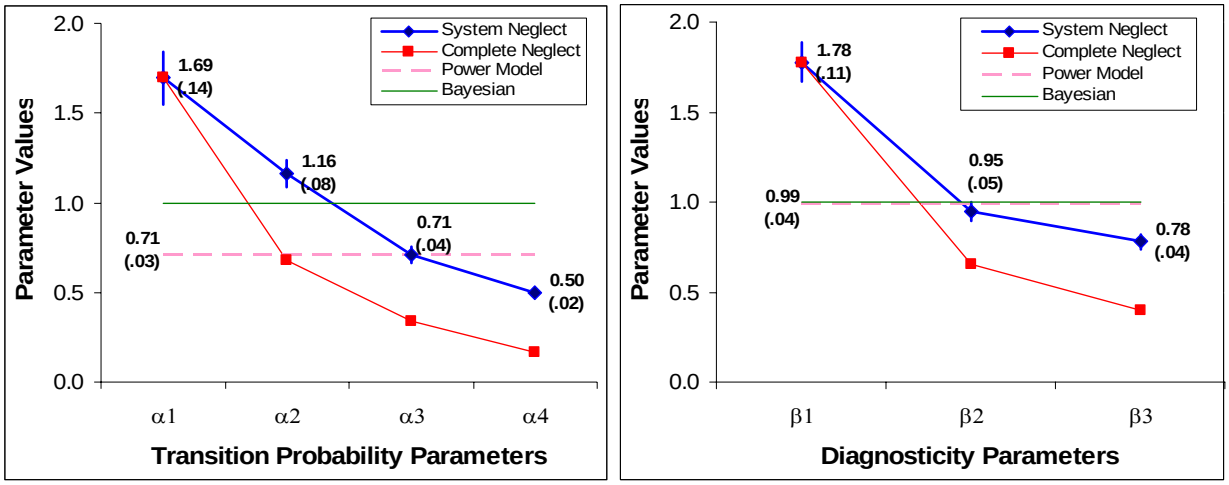


Figure 4: Parameter estimates for models fitting Study 1 median data using nonlinear regression. The left panel depicts the α parameters (transition probability), and the right panel depicts the β parameters (diagnosticity). Standard errors for the system neglect and power models are in parentheses and indicated by the vertical bars. A version of the complete neglect model is given as a theoretical baseline. Note that the complete-neglect model is not unique and includes all parallel translations.

Discussion

Study 1 reflects system neglect at two levels. First, subjects show the hypothesized gradient of under- and over-reaction: under-reaction is most prevalent in high diagnosticity/high transition probability systems, and over-reaction occurs most in low diagnosticity/low transition probability systems. Second, this pattern is reflected in our formal test of system neglect. Estimation of our quasi-Bayesian model produced diagnosticity and transition-probability parameters that were all ordered consistently with the system-neglect hypothesis. These estimates indicate that individuals are sensitive to changes in normatively relevant environmental parameters, but insufficiently so. Indeed, the parameter estimates are much closer to complete neglect than Bayesian updating. These findings echo the observations of previous researchers in non-stationary environments (Chinnis &

[5] We also estimated a "decay" model in which more distant observations received less weight. Although this model revealed reliable decay, the magnitude was slight and incorporating this additional parameter did not

Peterson, 1968; Barry & Pitz, 1979). Individual-level analyses provide similar support. Though there is significant heterogeneity, the vast majority of subjects exhibit system neglect.

On the whole, although we see slightly more under-reaction than over-reaction, we reiterate that the system-neglect hypothesis is agnostic about the overall level of under- and over-reaction. Rather, system neglect predicts a *relative* effect, not an absolute one.

4. STUDY 2: JUDGMENT AND ERROR

It is possible that the behavior we observe in Study 1 is a statistical artifact generated by regression effects or asymmetric error (Budescu, Erev, & Wallsten, 1997). For example, if the Bayesian standard calls for an estimate of 1, the only possible error is under-estimation. Erev *et al.* show how an error model can explain the co-existence of under- and over-confidence. Such error models are increasingly popular in both psychology and economics, and have been used to explain non-expected utility behavior (Hey & Orme, 1994; Hey, 1995; Ballinger & Wilcox, 1997), as well as overconfidence (Erev, Wallsten, & Budescu, 1994; Brenner, 2000). While there is little doubt that asymmetric error contributes to under- and over-reaction in both experimental and real-world settings, we believe that error is not necessary for producing the patterns observed in Study 1. To investigate this, we conducted a second judgment study to test specifically for the necessity of the error explanation.

In Study 2 we matched sequences to systems to produce the same Bayesian posterior. In some cases the system was strongly suggestive of change (*e.g.*, an unstable and precise system) while the signal was not, while in other cases the signal was strongly suggestive of change while the system was not (*e.g.*, a stable and noisy system). This design allows us to differentiate the system-neglect hypothesis from an error explanation. An error model of the type suggested by Erev *et al.* would predict no differences in reaction across the systems since we hold the Bayesian posterior constant. On the other hand, system neglect predicts a gradient identical to the one shown in Study 1, in which suggestive signals are given much more weight than suggestive systems.

---

significantly improve the model.

Methodology

The methodology used is identical to that used in Study 1 with a few notable exceptions. The experiment consisted of 2 practice trials followed by 30 trials, each consisting of 6 periods.

Our study used a 3 × 3 × 3 design, crossing three different transition probability levels (.05, .10, and .15), three different diagnosticity levels ($d = p_R / p_B$ of 1.5, 2.33, and 4), and three Bayesian posterior probabilities (.4, .5, and .6). We constructed 3 sequences for each of the 9 system cells to correspond to each Bayesian posterior probability level. Thus, sequences were matched to systems to produce nearly identical posterior probabilities. For example, sequences of $(r_1 = 1, r_2 = 1, r_3 = 1, r_4 = 0, r_5 = 0, r_6 = 0)$ and $(r_1 = 1, r_2 = 1, r_3 = 0, r_4 = 1, r_5 = 0, r_6 = 1)$ yielded posteriors of .413 and .415 for $(d = 1.5, q = .05)$ and $(d = 4.0, q = .15)$, respectively. The matches were approximate, ranging from posteriors of .381 to .420, .470 to .520, and .586 to .619. The average across all 3 matches ranged from .491 to .511 across the 9 cells. We also included 3 additional "filler" sequences to yield a total of 30 sequences. Each of the subjects received all 30 sequences in a randomized order. The actual sequences are found in ECP.

The need to match Bayesian posteriors across different systems influenced our decision to use sequences of 6 observations (as opposed to 10), as well as more moderate parameter values (*e.g.*, maximum diagnosticity of 4) and Bayesian posteriors (40-60%). Since the sequences were not drawn randomly, we also matched the frequency of the "actual regimes" to the Bayesian posteriors. For example, of the 15 instances in which the Bayesian posterior was between .45 and .55, we let 8 of 15 (53%) be generated by the blue bin and 7 of 15 (47%) be generated by the red bin.

We recruited 32 University of Chicago undergraduates as subjects. The scoring scheme was similar to that used in Study 1. We used a quadratic scoring system that paid $0.08 maximum (*e.g.*, if a subject indicated with certainty that the process was in the blue regime, and the process was in fact in the blue regime) and -$0.08 minimum (*e.g.*, if a subject indicated with certainty that the process was in the blue regime, and the process was in fact in the red regime).

Subjects were given feedback at the end of each trial as to if and when the process shifted from the red regime to the blue regime. They were also informed how much money they made (or lost) on that particular trial.

Results

We present both aggregate and individual-level results. Recall that our experiment required that each of our 32 subjects provide subjective probabilities that the process had switched to the blue regime for each of 6 signals in 30 trials. Let p_t^e be the empirical judgment and $| p_t^e - p_t^b |$ be the absolute difference between the subject's judgments and the Bayesian probability for signal *t*. Payments were based on the difference between their subjective probability of a change to the blue regime and the actual regime (1 if blue, 0 if red). The mean of this absolute difference was .33 (median=.20, sd=.35), generating an average payment of $7.70 (range of $4.80 to $9.35). The average payment to a Bayesian agent would have been $10.16.

To test for system neglect, we consider how our measure of under-reaction, $p_6^e - p_6^b$, changes across the 9 experimental cells. We consider only the 6th observation since posteriors are matched only on this observation. We show this measure pooled across the 3 posterior probability levels in Figure 5. We see the predicted gradient: over-reaction tends to be strongest when diagnosticity and transition probability are low, and under-reaction tends to be strongest when diagnosticity and transition probability are high. In pairwise comparisons of cells, the pattern of under-reaction is generally as predicted, with pairwise comparisons correctly ordered in 23 of 27 comparisons.

We also performed a similar individual-level analysis of the data. For each subject, we considered the pattern of under- and over-reaction, averaging the measure of reaction, $p_6^e - p_6^b$, over each of the 3 Bayesian posterior levels. We then counted how many of the 27 pairwise comparisons were ordered as required by system neglect. 29 of 32 subjects had a majority of comparisons in the predicted direction (14 or higher of 27). The number of pairwise comparisons in the predicted direction ranged from 5 to 27 (median=19, mean=18.6).

Estimation

We fit the system-neglect model to the Study 2 data using the same basic procedure outlined earlier. As before, we fit p_t^e, using nonlinear regression and the model in (3.3). This model was estimated for the 162 observations using the median judgment for each observation. The estimates are found in Figure 6. The pattern of parameter estimates looks very similar to that found with Study 1. The slope of the estimates is consistent with system neglect, and the parameter estimates are close to what would be obtained by a complete-neglect model.

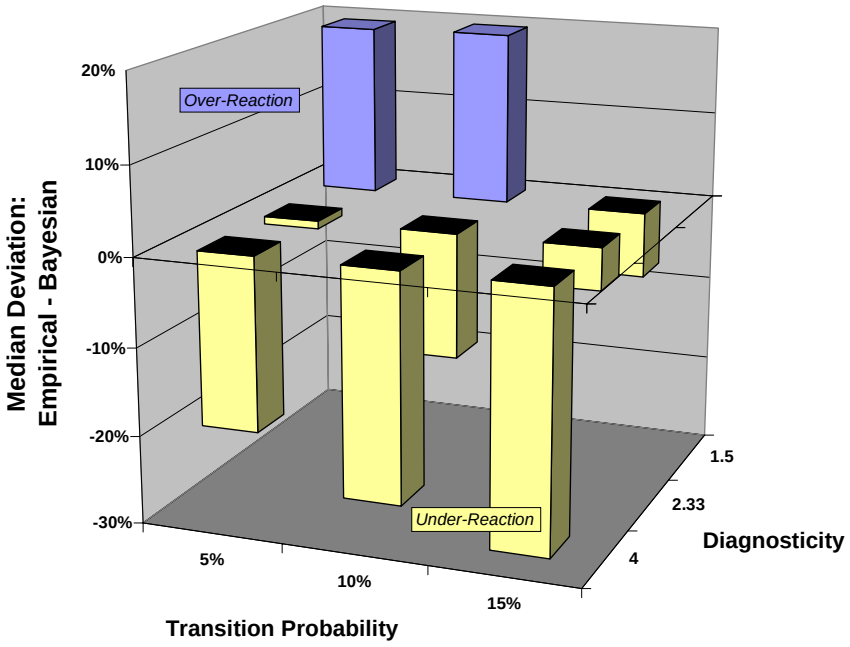


Figure 5: Over- and under-reaction, by condition, for each of the 3 levels of posterior probability and aggregated over the 3 levels, as measured by the difference between median empirical probability judgments and the median Bayesian probabilities, $p_6^e - p_6^b$ (Study 2)

Discussion

Study 2 investigated the possible role of random error in the pattern found in Study 1. Whereas asymmetric error almost certainly contributes to system neglect, Study 2 indicates there is more to system neglect than *just* error. We find the same strong pattern of under- and over-reaction predicted by the system-neglect hypothesis even after neutralizing the role of error. This study also demonstrates the extent to which signals are privileged over systems. By design, the gradient we observe in this study reveals the relative emphasis of signals over systems.

Unlike Study 1, the sequences used in Study 2 were not randomly generated. We did so to create a very stark test between the error account and system neglect. In theory, this approach could bias judgments in favor of the system-neglect hypothesis, as subjects may discount system parameters after observing a sequence that is not representative of a particular system. We regard this possibility as highly unlikely since it would require that subjects are sensitive to both the underlying systems and the representativeness of sequences drawn from them. Overall, we believe the approach in this study complements Study 1, in which we use randomly generated sequences.

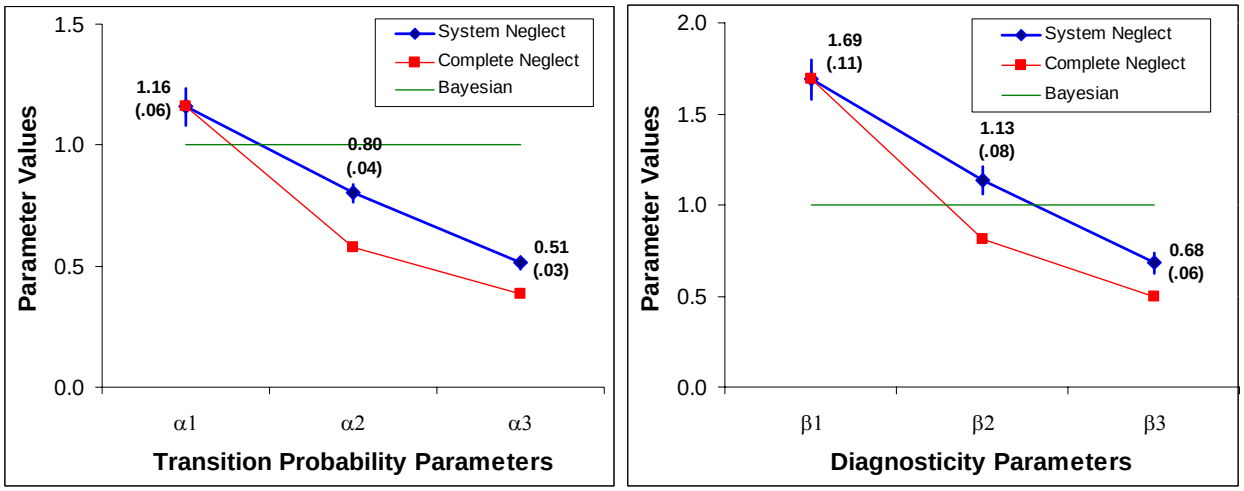


Figure 6: Parameter estimates for models fitting Study 2 median data using nonlinear regression. The left panel depicts the α parameters (transition probability), and the right panel depicts the β parameters (diagnosticity). Standard errors for the system-neglect and power models are in parentheses and indicated by the vertical bars. A version of the complete-neglect model is given as a theoretical baseline. Note that the complete-neglect model is not unique and includes all parallel translations.

5. Study 3: Choice Task

The experimental design of Study 3 was similar to that of Studies 1 and 2. The major difference was that subjects were asked to predict the color of the next ball rather than to provide a probability estimate.

Experimental Design

The program, written in Visual Basic, included several screens of introduction, 2 practice trials, and 18 trials of 10 periods each. For each trial, subjects were shown the relevant parameters, p_R, p_B, and q, governing that trial. They were then shown a sequence of red or blue balls drawn

randomly based on the set of parameters. Before seeing each signal, subjects were asked to predict the color of the next ball.

As in Study 1, we varied diagnosticity and transition probability. Study 3 used almost the same parameters as Study 1: $(p_R, p_B) = (.6,.4)$, $(.75,.25)$, and $(.9,.1)$ and $q = .025$, .05, .10, and .20 ($q = .02$ was used in Study 1). Thus our 3×4 design yields 12 experimental conditions. We randomly generated 3 unique sequences for each of these 12 conditions, creating 36 total sequences. Each subject experienced half of the sequences, and either 1 or 2 of the 3 sequences from each condition. Sequences were randomized for each subject.

We recruited 50 University of Chicago students to participate in this study. The median number of undergraduate and graduate mathematics and statistics classes taken by our subjects was again 3. Subjects did not receive feedback regarding if or when a change actually occurred, nor were they shown the optimal (Bayesian) responses. They did, however, find out immediately whether their prediction was correct. We paid subjects 9 cents for each correct prediction (out of 180).

Normative Model

As in Study 1, we use the Bayesian response as the standard for evaluating subject behavior. Recall that (3.1) gives us the Bayesian probability that the process has switched to the blue regime at $t-1$: $p^b_{t-1} = \Pr(B_{t-1} \mid H_{t-1})$. The probability, then, that a ball is drawn from the blue regime at *t*, $\Pr(B_t \mid H_{t-1})$, is $p^b_{t-1} + (1 - p^b_{t-1})q$, and the probability that a blue ball is observed, $\Pr(r_t = 0 \mid H_{t-1})$, is $(p^b_{t-1} + (1 - p^b_{t-1})q)(1 - p_B) + (1 - p^b_{t-1})(1 - q)(1 - p_R)$, which is greater than .5 if and only if $\Pr(B_t \mid H_{t-1}) = p^b_{t-1} + (1 - p^b_{t-1})q > .5$. Normatively, then, a subject should predict a blue ball if she believes the process is in the blue regime, and predict a red ball otherwise (assuming the rewards for correct predictions are symmetric, as is true in our study).

Experimental Results

We begin by comparing the accuracy of the empirical predictions to the accuracy of the Bayesian predictions. The Bayesian predictions (69%) were better than the empirical predictions (64%). However, there was considerable heterogeneity in accuracy across subjects, with accuracy

rates ranging from 59% to 73%. Thus, payments ranged from $9.63 to $11.79 (mean=$10.62). A Bayesian agent would make $11.20 on average.

As in Study 1, our primary interest is belief revision. In the present context, we observe belief revisions when subjects change their predictions. Let c_t^e (c_t^b) be the subject's prediction (Bayesian prediction) of the color of the *t*-th ball ($c_t = 0$ if a blue ball is predicted). Thus, an *empirical belief-revision* corresponds to $c_t^e \neq c_{t-1}^e$, whereas a *Bayesian belief-revision* corresponds to $c_t^b \neq c_{t-1}^b$. Fewer empirical than Bayesian belief-revisions implies under-reaction, while the opposite implies over-reaction. Across all conditions, we observe belief revisions at the rate of 16.1% (empirical) and 11.3% (Bayesian). Thus, on average, subjects revise their predictions 42% more often than the normative prediction. By this measure, there is an overall tendency to over-react.

We next turn to how this measure varies across experimental conditions. Recall that system neglect implies over-reaction will be most extreme in stable/noisy environments and least extreme in unstable/precise environments. Figure 7 depicts the difference between the percentage of empirical and Bayesian belief-revisions across all 12 experimental conditions. Positive values indicate that empirical changes exceed Bayesian changes, and, thus, over-reaction. We observe the predicted gradient between the southeast cell (unstable/precise) and the northwest cell (stable/noisy). Under-reaction is concentrated in the southeast corner and over-reaction in the northwest corner, as system neglect suggests.

As an alternative measure of over- and under-reaction, we consider the amount of evidence required to induce a blue-ball prediction. Normatively, this threshold should not vary across conditions: as soon as the data suggest that the blue regime is more likely than the red regime, subjects should predict a blue ball. The system-neglect hypothesis, on the other hand, suggests that individuals will require substantially more evidence, objectively measured, in unstable/precise environments than in stable/noisy environments.

To test this prediction, we analyzed behavior in the following way. For every trial that a subject completed, we identified the period in which she first predicted a blue ball. This period indicates the point at which the subject believes a regime shift has occurred. We then calculate the

normative probability underlying that prediction, *i.e.*, the Bayesian prior that a blue ball will be drawn in that period. This is the evidence, *objectively* measured, on which the subject bases her prediction. Finally, we take the mean of these probabilities for each experimental condition, pooling all trials and subjects within each condition. Figure 8 compares these means across all 12 experimental conditions.

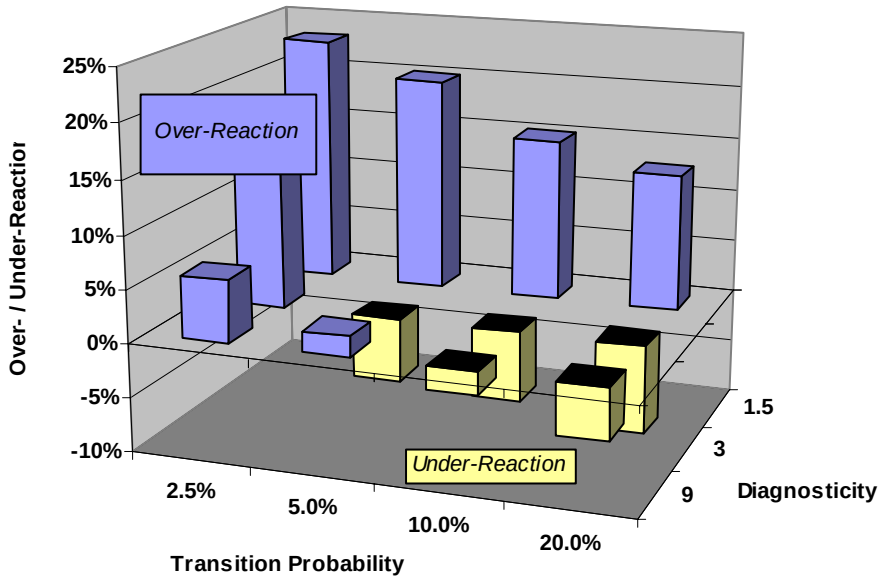


Figure 7: Over- and under-reaction as measured by belief revisions (Study 3, Choice Task). Plotted are the differences in the frequency of Empirical and Bayesian belief-revisions for each condition.

As predicted, thresholds are considerably higher in the southeast corner than in the northwest. On average, subjects in the most unstable/precise environment required evidence indicating a .90 chance that the system has shifted to the blue regime in order to make their first blue prediction, *i.e.*, they under-reacted. Conversely, when in the most stable/noisy environment they required only a .11 chance that the system has shifted to the blue regime in order to make their first blue prediction, *i.e.*, they over-reacted.

Estimation

To formally test the system-neglect hypothesis for our choice task, we adapt our estimation procedure to account for the binary dependent variable. We take a "single-agent stochastic choice model" approach (*e.g.*, Camerer & Ho, 1994; Wu & Gonzalez, 1996), modeling a "representative

agent" with noise (see also, McFadden, 1981). There are two steps to this process. In the first step, we obtain a subjective probability of drawing the next ball from the blue regime. We assume subjects base their prediction on this probability. In the second step, we transform this subjective probability into a stochastic prediction about the next observation, *i.e.*, the probability that subjects will predict a blue ball. The objective of this model is to match this probability with the observed percentage of subjects predicting a blue ball.

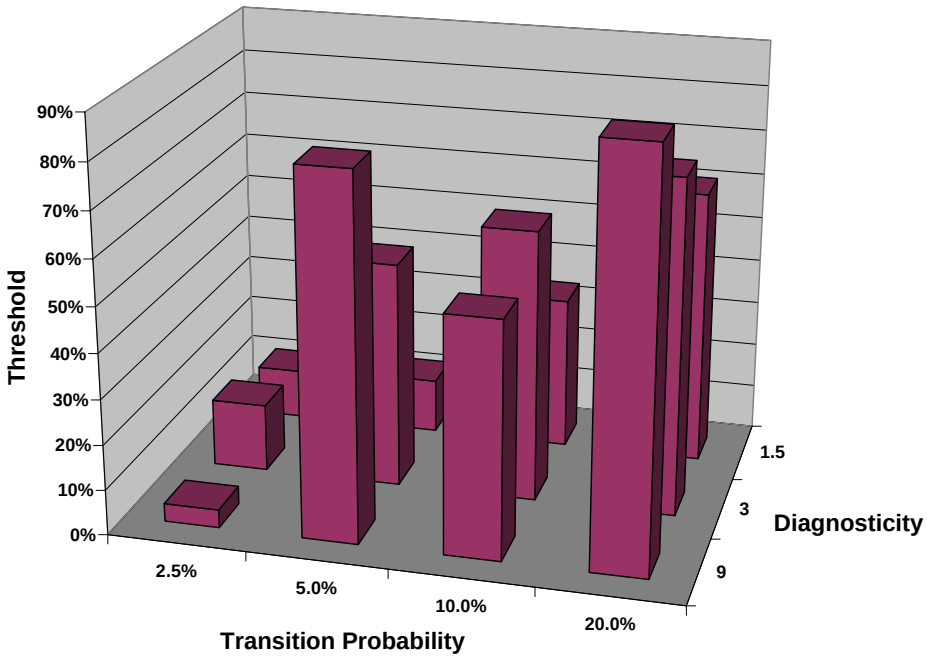


Figure 8: Mean Bayesian probability of having changed to the blue regime at the time of the first blue prediction (Study 3, Choice Task). Bayesian probabilities are calculated at the time of the first blue-ball prediction by each subject in each trial. This summary pools subjects and trials within each condition.

Recall that p^e_{t-1} denotes the subjective probability that the $t-1$-th ball was drawn from the blue regime. Since our subjects predict the *t*-th ball, we need to advance p^e_{t-1} one period as we did with the normative model. Therefore, we take $\pi^e_t = (p^e_{t-1} + (1-p^e_{t-1})q)(1-p_B) + (1-p^e_{t-1})(1-q)(1-p_R)$ to be the subjective probability that a blue ball will be drawn from the blue regime at *t*. Note that this is simply the normative expression with p^e_{t-1} substituted for p^b_{t-1}.

Next we fit this subjective probability using (3.3) to the observed predictions. We denote the percentage of subjects who predicted a blue ball for ball *t*, $\%B_t$. In the normative model, subjects

would predict a blue ball if $\pi_t^e > .5$ (thus, $\%B_{it}$ would always be either 0 or 1). To fit empirical observations, we assume a stochastic choice functional which allows the representative agent to be imperfectly sensitive to π_t^e,

$$1/\left(1+\exp\left[\eta-\psi\pi_t^e\right]\right), \tag{5.1}$$

where π_t^e is estimated, as detailed in the previous paragraph, using either the power model or the system-neglect model. Note that the response function captures how discriminating the pattern of predictions is to changes in π_t^e with discriminability increasing in η, and that (5.1) is symmetric if $\eta = .5\psi$. We estimate η and ψ (as well as the parameters needed to fit π_t^e) using nonlinear least-squares. We take the "error" to be the difference between (5.1) and $\%B_t$, the percentage of subjects who predicted a blue ball, and choose the parameters that minimize the sum of squared errors.

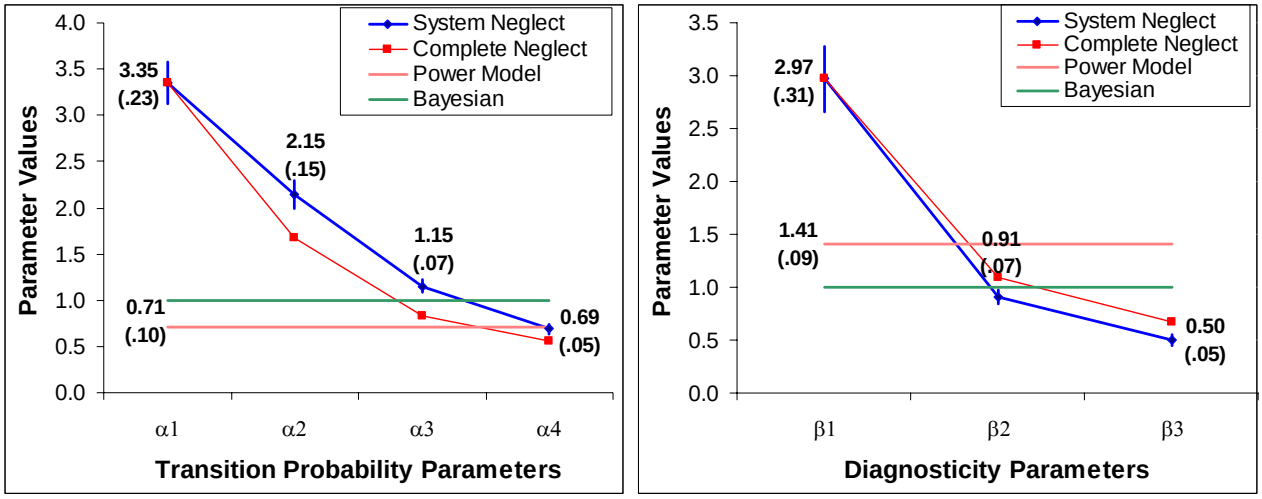


Figure 9: Parameter estimates for models fitting Study 3 using the stochastic choice functional. The left panel depicts the α parameters (transition probability), and the right panel depicts the β parameters (diagnosticity). Standard errors for the system-neglect and power models are in parentheses and indicated by the vertical bars. A version of the complete-neglect model is given as a theoretical baseline. Note that the complete-neglect model is not unique and includes all parallel translations.

The estimates for the power model and the system-neglect model are found in Figure 9. Most significantly, model estimates for both parameters are ordered in the manner predicted by the system-neglect hypothesis. As in Study 1, we take this ordering as the formal test of the system-neglect hypothesis. Recall that system neglect predicts that parameter estimates will decrease as

system parameters increase, which for the present study implies $\alpha_m > \alpha_n$ and $\beta_m > \beta_n$ for $m < n$. All 9 (6 α and 3 β) of the comparisons are ordered in the direction predicted by system neglect and highly significant ($p < .001$). The estimates also indicate significant variation in the degree of reaction across environments: for both parameters we find a mixed pattern of over-reaction ($\alpha_m, \beta_n > 1$ for low *m* and *n*) and under-reaction ($\alpha_m, \beta_n < 1$ for high *m* and *n*).[6]

Figure 9 also shows the estimates for the system-neglect model relative to the Bayesian, power, and complete-neglect models. Note that the system-neglect model is significantly steeper than either the Bayesian model ($\alpha_m = \beta_n = 1$) or the power model ($\alpha = .71, \beta = 1.41$), reflecting the importance of incorporating system neglect when modeling behavior in this domain. Note also that whereas the system-neglect model differs from the complete-neglect model for the α parameters, it is virtually indistinguishable from the complete-neglect model for the β parameters.

Discussion

In Study 3, we changed the task from judgment to choice. We again see evidence of system neglect at two levels, graphically and formally in our parameter estimates. We measure under- and over-reaction two different ways: switches in predictions or evidence required to make first blue prediction. Both measures show the posited pattern: the greatest under-reaction occurs in the stable/noisy conditions (southeast), and the greatest over-reaction occurs in the unstable/precise conditions (northwest). It is noteworthy that we observe the same general pattern even though the measure of reaction is very different in Study 3 than in Study 1 (change in probability judgments), and the task is in Study 3 is more natural and less cognitively taxing.

We also find formal support for system neglect. The system-neglect model gives a perfect ordering of the parameter estimates. In sum, as in Study 1, subjects are sensitive to changing environmental conditions, but insufficiently so.[7] The estimation method used in Study 3 differs from

[6] The response function is close to symmetric: $\eta = 7.58$ and $\psi = 14.22$ for the system-neglect model. The system-neglect model ($R^2 = .97$) fits the data better than either a power model ($R^2 = .83$) or a Bayesian model ($R^2 = .82$).

[7] We do not estimate individual-level parameter values in Study 3 because the binary nature of the dependent variable does not provide the statistical power to estimate individual-level models.

the method used in Study 1 in that the dependent variable in Study 3 is the percentage of blue ball predictions. Thus, we should be cautious in comparing the parameter estimates across the three studies. That said, the estimates for α-parameters are remarkably similar, while the β-parameters in Study 3 are more elevated than the β-parameters in Study 1, indicating a greater tendency to over-react in Study 3.

Recall that the system-neglect hypothesis addresses the *relative* likelihood of under- and over-reaction across two environments and hence is agnostic on the overall level of over-reaction in a particular environment. However, we believe psychological research in two other areas sheds light on why these differences might occur. First, anchoring and insufficient adjustment (Tversky & Kahneman, 1974) is more likely to occur in Study 1 than in Study 3. Subjects in Study 1 may anchor on the system's initial values (i.e., 0) or on their own previous estimates, then adjust too slowly in response to new data. This process is less pronounced in Study 3 because binary choices preclude small adjustments, as large discrete jumps are required. Second, stimuli-response compatibility is different in the two studies. In both studies subjects respond to binary stimuli, either red or blue balls. However, the manner in which subjects respond is different: with probability estimates in Study 1 and with choices in Study 3. Thus, stimuli and response are compatible (both binary) in Study 3 and incompatible in Study 1. Psychological research in other domains has emphasized the influence of this type of compatibility (Slovic, Griffin & Tversky, 1990; Fischer & Hawkins, 1993). Indeed, compatibility has been put forth as one explanation for the classic preference-reversal demonstrations (Slovic & Lichtenstein, 1968; Grether & Plott, 1979; Tversky, Slovic, & Kahneman, 1990). It may also be that stimuli-response compatibility leads to more over-reaction in Study 3 since it facilitates "chasing noise." Although a more complete psychological story awaits future research, together these two lines of research offer complementary explanations for the observed *levels* of responsiveness: anchoring facilitates under-reaction in Study 1, and compatibility facilitates over-reaction in Study 3. Again, however, absolute levels of responsiveness are not the focus of this paper.

5. CONCLUSION

We investigated the ability of individuals to detect and respond to changes in a dynamic environment. Normatively, behavior in these environments *should* depend on the system governing the dynamic process. We find that while subjects are sensitive to system parameters (namely, diagnosticity and transition probability), they are insufficiently so. We investigated system neglect using different tasks (a judgment task in studies 1 and 2 and a choice task in study 3) and tests (by evaluating the pattern of under- and over-reaction and by formally estimating parameter values across experimental conditions). In all cases we found strong support for the system-neglect hypothesis.

Contributions

Our findings extend research in both stationary and non-stationary environments. First, it adds to a growing body of psychological research that shows that individuals consistently overweight the strength of evidence (such as indications of change) at the expense of the weight of evidence (such as the system that produced those indications). The same psychological mechanism that reconciles conservatism and representativeness in stationary judgment tasks seems to explain our results. Koehler, Brenner, & Griffin (2002) have argued for a more general case-based approach to judgment: individuals evaluate the particular case at hand, with little regard to the class to which the case belongs. This attractive account unifies many of the classic findings in the heuristics and biases literature.

Second, our results contribute empirically and theoretically to our understanding of how individuals make judgments in nonstationary settings. Fischhoff & Beyth-Marom (1983) noted that research in conservatism "was quietly abandoned" partly because of the mixed results and failure to identify definitive psychological mechanisms. For many of the same reasons, research in regime-shift detection has been dormant for thirty years. Our studies find both under- and over-reaction, as was first demonstrated in the 1960's, but explain both types of responses in terms of a single

theoretical notion.[8] Our models have the additional advantage of providing a continuous parametric specification that ranges from complete neglect to Bayesian.

Third, research in financial economics has documented the tendency of investors to under-react to short-term earnings news and over-react to long-term earnings news. Griffin & Tversky's (1992) strength and weight account has motivated theoretical models designed to explain these findings (*e.g.*, Barberis *et al*. 1998; Daniel *et al*. 1998), as well as experimental financial economics investigations (*e.g.*, Nelson *et al.* 2001; Bloomfield *et al*. 2002). Indeed, Barberis *et al*. (1998) and Brav & Heaton (2002) have used regime-shift models to explain these regularities. In these models, investors' belief-revisions are Bayesian, an assumption challenged by the system-neglect hypothesis. Our research suggests a very specific way in which individuals are quasi-Bayesian, and thus calls for greater attention to the precision of market information about valuation-relevant parameters, as well as the stability of those parameters. Investor insensitivity to these "system" parameters may well be a culprit in the under- and over-reaction found in asset prices.

Alternative Explanations

We consider two possible alternative explanations for our results, risk aversion and random error. In Study 1, we used a quadratic scoring rule that is truth-revealing for risk-neutral subjects. Risk-neutrality is commonly assumed for the stakes involved in this study (Davis & Holt, 1993). However, the quadratic scoring rule has a "flat maximum", thus shading stated probabilities toward .5 reduces variance substantially while sacrificing little expected value. Consequently, risk-aversion, if sufficiently extreme, could play a role in the pattern demonstrated in Study 1. However, it would not predict the pattern found in Study 2, where we held the Bayesian posterior probability constant. It also plays no role in Study 3, since optimal behavior relies only on monotonic subjective probabilities. Thus, risk-aversion cannot provide a general explanation for all experimental results. Moreover, we have found results similar to those in Study 1 using linear payoff schemes that do not have a "flat maximum".

[8] Indeed, Edwards (1968) actually reports evidence contrary to conservatism. His Figure 2.2 reports accuracy ratios for three diagnosticity levels (accuracy ratios less than 1 indicate conservatism, while those greater than 1 indicate

Likewise, error cannot provide a complete account of our pattern of data. Erev *et al*. offer an error story in which responses are unbiased "true judgments" perturbed by error. This account requires that the median response be unbiased. As expected, our pattern of over- and under-reaction is less pronounced in median than mean data, but still highly significant. Study 2 deals with the error account directly. Consistent with system neglect, we found a gradient identical to the one shown in Study 1: subjects believed that the chance of a change was significantly higher when given a strong signal in a weak system than when given a weak signal in a strong system.

Future Directions

Our evidence for system neglect thus far is limited to an experimental framework with symmetric probabilities, binary signals, and absorbing states. Although we expect these findings to be robust, system neglect in alternative environments merits direct investigation. We have communicated the true parameters of the statistical processes to our subjects, but in most situations a decision maker is not aware of these parameters. The question of how individuals identify the relevant parameters of an unspecified process is an important question that deserves further investigation. It is also important to know whether experience improves performance, *i.e.*, whether subjects can learn to detect regime-shifts more effectively. We have conducted judgment and choice studies in which subjects remain in one system for the entire study. These preliminary learning studies indicate that experience can improve performance, but that this improvement is slight and does not occur in all conditions (Wu & Massey, 2003).

While the system-neglect hypothesis is agnostic about levels of under- or over-reaction, a number of psychological factors might influence the responsiveness to signals in the laboratory and real world. We have found that the decision task, judgment versus choice, influences the level of under- or over-reaction. An additional factor that deserves attention is motivated reasoning. Kunda (1990) has argued that individuals are motivated processors of ambiguous evidence. Thus, a decision maker with a vested interest in the continuation of an incumbent regime (such as a "bricks and

---

radical belief-revision). The accuracy ratios differ in the direction system neglect would predict, with the lowest diagnosticity condition (d=.55/.45) actually showing *radical* belief-revision.

mortar" retailer) might be slower to react than someone who profits from a new regime (such as an "e-tailer").

In addition, the payoffs in most business situations are not symmetric, *i.e.*, under-reaction and over-reaction are not punished equally. Rather, changing strategies is costly in real terms, whether it be shifting resources, building or closing new facilities, or acquiring new expertise. These costs may provide an economic rationalization, legitimate or not, for under-reaction. Moreover, the susceptibility of decision makers to sunk costs and commitment escalation creates a psychological asymmetry as well, which can lead decision makers to believe too long in the current regime (*e.g.*, Thaler, 1980; Staw, 1981).

We conclude by discussing some implications for decision making. Much popular management literature has called for organizations to be nimble, reactive, and flexible (Bhide, 2000; Hamel, 2000; Schoemaker, 1995). What often goes unsaid is that organizations must be *appropriately* reactive. Clearly there is a trade-off between flexibility and perseverance. This research suggests there are environments in which decision makers are prone to stay the course when they should change direction, and change direction when they should stay the course.

Dixit & Nalebuff (1991) suggest, "Even though you can't guess right all the time, you can at least recognize the odds." (p.169) We push that sentiment back a step: "Even if you can't calculate the odds, you can at least understand your environment." Our research shows that understanding one's environment is critical to reacting appropriately. Indeed, regardless of whether the task is to "guess right" or to "calculate the odds", individuals emphasize indications-of-change over the environment producing those indications. Decision makers would be well served to counter this bias by allocating more resources to evaluating their environment, actively investigating signal diagnosticity and system stability. Understanding one's environment is crucial to correctly interpreting events and, hence, managing the tension between under- and over-reaction.

APPENDIX: DERIVATION OF BAYESIAN RESPONSE

Consider a Bernoulli process that can take on one of two regimes, a "red" probability regime (*R*) and a "blue" probability regime (*B*). Let θ_t be the regime in period *t*. R_t (B_t) indicates that the process is in the red (blue) regime in period *t*, *i.e.*, $\theta_t = R$ (and $\theta_t = B$). The process begins in the red regime (*i.e.*, $\theta_0 = R$ or R_0). However, the process may switch to the blue regime at any time including period 1. The probability of the process changing from the red regime to the blue regime is given by a "transition probability", $q = \Pr(B_t \mid R_{t-1})$. We assume that the blue regime is an absorbing state, $\Pr(B_t \mid B_{t-1}) = 1$.

Let r_t be an indicator variable, such that $r_t = 1$ if a red ball is drawn and $r_t = 0$ otherwise, and $H_t = (r_1, ..., r_t)$ be the history of signals. Finally, let $p_R = \Pr(r_t = 1 \mid R_t)$ and $p_B = \Pr(r_t = 1 \mid B_t)$ for all *t* denote the probability of a red signal in the Red and Blue regimes respectively, and $d = p_R / p_B$ denote diagnosticity.

To calculate the posterior odds, $\dfrac{\Pr(B_t \mid H_t)}{\Pr(R_t \mid H_t)}$, we consider the numerator and denominator separately:

$$\Pr(B_t \mid H_t) = \sum_{j=1}^{t} \left(\prod_{k=1}^{j-1} \Pr(R_k \mid R_{k-1}) \Pr(r_k \mid R_k) \right) \left(\prod_{k=j}^{t} \Pr(B_j \mid R_{j-1}) \Pr(r_k \mid B_k) \right) =$$

$$\sum_{j=1}^{t} q(1-q)^{j-1} \prod_{k=1}^{j-1} \Pr(r_k \mid R_k) \prod_{k=j}^{t} \Pr(r_k \mid B_k), \tag{1}$$

and

$$\Pr(R_t \mid H_t) = \prod_{k=1}^{t} \Pr(R_k \mid R_{k-1}) \Pr(r_k \mid R_k) = (1-q)^t \prod_{k=1}^{t} \Pr(r_k \mid R_k). \tag{2}$$

Note that (1) captures the multiple paths that might result in $\theta_t = B$, *i.e.*, $\{\theta_1 = B\}$, $\{\theta_1 = R, \theta_2 = B\}, ..., \{\theta_1 = R, ..., \theta_{t-1} = R, \theta_t = B\}$. Using (1) and (2), we write the posterior odds as

$$\frac{\Pr(B_t \mid H_t)}{\Pr(R_t \mid H_t)} = \frac{\sum_{j=1}^{t} q(1-q)^{j-1} \prod_{k=1}^{j-1} \Pr(r_k \mid R_k) \prod_{k=j}^{t} \Pr(r_k \mid B_k)}{(1-q)^t \prod_{k=1}^{t} \Pr(r_k \mid R_k)} = \sum_{j=1}^{t} \frac{q}{(1-q)^{t-j+1}} \prod_{k=j}^{t} \frac{\Pr(r_k \mid B_k)}{\Pr(r_k \mid R_k)}. \tag{3}$$

Noting that $\dfrac{\Pr(r_k \mid B_k)}{\Pr(r_k \mid R_k)} = \left(\dfrac{p_B}{p_R} \right)^{r_k} \left(\dfrac{1-p_B}{1-p_R} \right)^{1-r_k}$ and, for the special case of a symmetric process, $p_R = 1 - p_B$, $\dfrac{P(r_j \mid B)}{P(r_j \mid R)} = \left(\dfrac{p_B}{p_R} \right)^{2r_j - 1} = d^{1-2r_j}$, we can simplify (3) to

$$\frac{\Pr(B_t \mid H_t)}{\Pr(R_t \mid H_t)} = \sum_{j=1}^{t} \frac{q}{(1-q)^{t-j+1}} \prod_{k=j}^{t} d^{1-2r_k} = \sum_{j=1}^{t} \frac{q}{(1-q)^{t-j+1}} d^{\,t-j+1-\left(2\sum_{k=j}^{t} r_k\right)}. \tag{4}$$

The exponent on d is simply the difference between the number of blues and number of reds over the path being evaluated.

Note that the unconditional likelihood is given by

$$\frac{\Pr(B_t)}{\Pr(R_t)} = \frac{1-(1-q)^t}{(1-q)^t}, \tag{5}$$

which expresses the "base rate" odds of being in B_t and R_t. Factoring (5) out of (4), we get

$$\frac{\Pr(B_t \mid H_t)}{\Pr(R_t \mid H_t)} = \left(\frac{1-(1-q)^t}{(1-q)^t}\right) \sum_{j=1}^{t} \frac{q(1-q)^{j-1}}{1-(1-q)^t} d^{\,t+1-j-\left(2\sum_{k=j}^{t} r_k\right)}. \tag{6}$$

REFERENCES

Ballinger, T. Parker and Nathanial T. Wilcox (1997). "Decisions, error and heterogeneity." *Economic Journal* 106, 1090-1105.

Barry, Donald M. and Gordon F. Pitz (1979). "Detection of Change in Nonstationary, Random Sequences." *Organizational Behavior and Human Performance* 24, 111-125.

Bhide, Amar (2000). *The Origin and Evolution of New Businesses*. New York: Oxford University Press.

Bloomfield, Robert and Jeffrey Hales (2002). "Predicting the next step of a random walk: experimental evidence of regime-shifting beliefs." *Journal of Financial Economics* 65, 397-414.

Brenner, Lyle (2000). "Should observed overconfidence be dismissed as a statistical artifact?: Critique of Erev, Wallsten, and Budescu (1994)." *Psychological Review* 107, 943-946.

Brier, Glenn W. (1950). "Verification of forecasts expressed in terms of probability." *Monthly Weather Review* 78, 1-3.

Brown, Eric R. and Alice L. Bane (1975). "Probability estimation in a chance task with changing probabilities." *Journal of Experimental Psychology: Human Perception & Performance* 1, 183-187.

Budescu, David V., Ido Erev and Thomas S. Wallsten (1997). "On the Importance of Random Error in the Study of Probability Judgment Part I: New Theoretical Developments." *Journal of Behavioral Decision Making* 10, 157-171.

Camerer, Colin F. and Teck-Hua Ho (1994). "Violations of the Betweenness Axiom and Nonlinearity in Probability." *Journal of Risk and Uncertainty* 8, 167-196.

Chinnis, James O. and Cameron R. Peterson (1968). "Inference about a Nonstationary Process." *Journal of Experimental Psychology* 77, 620-625.

Chinnis, James O. and Cameron R. Peterson (1970). "Nonstationary Processes and Conservative Inference." *Journal of Experimental Psychology* 84, 248-251.

Crown, Judith and Glenn Coleman (1996). *No Hands: The Rise and Fall of the Schwinn Bicycle Company, an American Institution*. New York: Henry Holt.

Daniel, Kent, David Hirshleifer and Avandihar Subrahmanyam (1988). "Investor psychology and security market under- and overreactions." *Journal of Finance* 53, 1839-1885.

Davis, Douglas D. and Charles A. Holt (1993). *Experimental Economics*. Princeton, NJ, Princeton Unversity Press.

Dixit, Avinash K., and Barry J. Nalebuff (1991). *Thinking Strategically: The Competitive Edge in Business, Politics, and Everyday Life*, New York: W.W. Norton & Company.

Edwards, Ward (1968). "Conservatism in human information processing." In *Formal representation of human judgment*. Benjamin Kleinmuntz (Ed.). New York: Wiley, 17-52.

Erev, Ido, Thomas S. Wallsten and David V. Budescu (1994). "Simultaneous over- and underconfidence: the role of error in judgment processes." *Psychological Review* 101, 519-527.

Estes, William K. (1954). "Individual Behavior in Uncertain Situations: An Interpretation in Terms of Statistical Association Theory." In *Decision Processes*. Robert M. Thrall, Clyde H. Coombs and Robert L. Davis (Eds.). New York: John Wiley, 127-138.

Estes, William K. (1984). "Global and local control of choice behavior by cyclically varying outcome probabilities." *Journal of Experiment Psychology: Learning, Memory, and Cognition* 10, 258-270.

Fischhoff, Baruch and Ruth Beyth-Marom (1983). "Hypothesis evaluation from a Bayesian perspective." *Psychological Review* 90, 239-260.

Fischer, Gregory W. and Scott A. Hawkins (1993). "Strategy compatibility, scale compatibility, and the prominence effect." *Journal of Experimental Psychology: Human Perception and Performance* 46, 835-847.

Flood, Merrill H. (1954). "Environmental Non-Stationarity in a sequential decision-making experiment." In *Decision Processes*. Robert M. Thrall, Clyde H. Coombs and Robert L. Davis (Eds.). New York: John Wiley, 139-158.

Grether, David M. (1980). "Bayes Rule as a descriptive model: The representativeness heuristic." *Quarterly Journal of Economics* 95, 537-557.

Grether, David M. and Charles R. Plott (1979). "Economic theory of choice and the preference reversal phenomenon." *American Economic Review* 69, 623-638.

Griffin, Dale and Amos Tversky (1992). "The Weighing of Evidence and the Determinants of Confidence." *Cognitive Psychology* 24, 411-435.

Grove, Andrew S. (1999). *Only the Paranoid Survive: How to Exploit the Crisis Points that Challenge Every Company*. New York: Doubleday.

Hamel, Gary. (2000). *Leading the Revolution*. Boston, MA: Harvard Business School Press.

Jones, Edward E., & Harris, Victor A. (1967). "The attribution of attitudes." *Journal of Experimental Social Psychology* 3, 1-24.

Kahneman, Daniel and Amos Tversky (1973). "Subjective probability: A judgment of representativeness." *Cognitive Psychology* 3, 430-454.

Koehler, Derek J., Lyle Brenner and Dale Griffin (2002). "The Calibration of Expert Judgment: Heuristics and Biases Beyond the Laboratory." In *Heuristics and Biases: The Psychology of Human Judgment*. Thomas Gilovich, Dale Griffin and Daniel Kahneman (Eds.). Cambridge, England, Cambridge University Press, 686-715.

McFadden, Daniel. (1981). "Econometric Models of Probabilistic Choice," *Structural Analysis of Discrete Data with Econometric Applications.* Charles F. Manski and Daniel McFadden (Eds.). Cambridge: MIT Press, 198-272.

Nelson, Mark W., Robert Bloomfield, Jeffrey W. Hales and Robert Libby (2001). "The Effect of Information Strength and Weight on Behavior in Financial Markets." *Organizational Behavior & Human Decision Processes* 86, 168-196.

Rapoport, Amnon, William E. Stein and Graham J. Burkheimer (1979). *Response models for detection of change*. Dordrecht, Holland: D. Reidel.

Robinson, Gordon H. (1964). "Continuous estimation of a time-varying probability." *Ergonomics* 7, 7-21.

Schoemaker, Paul J.H. (1995). "Scenario Planning: A Tool for Strategic Thinking." *Sloan Management Review* (Winter), 25-40.

Slovic, Paul, Dale Griffin, and Amos Tversky (1990). "Compatibility effects in judgment and choice." *Insights in Decision Making: A Tribute to Hillel J. Einhorn*. Robin Hogarth (Ed.). Chicago: Chicago University Press, 5-27.

Slovic, Paul and Sarah Lichtenstein (1968). "The relative importance of probabilities and payoffs in risk taking." *Journal of Experimental Psychology* 78, 1-18.

Smith, A.F.M. (1975). "A Bayesian Approach to Inference About a Change-Point in a Sequence of Random Variables." *Biometrika* 62, 407-416.

Staw, Barry M. (1981). "The escalation of commitment to a course of action." *Academy of Management Journal* 6, 577-587.

Theios, John, John W. Brelsford and Phyllis Ryan (1971). "Detection of change in nonstationary binary sequences." *Perception and Psychophysics* 9, 489-492.

Tversky, Amos and Daniel Kahneman (1974). "Judgment under Uncertainty: Heuristics and Biases." *Science* 184, 1124-31.

Thaler, Richard H (1980). "Toward a Positive Theory of Consumer Choice," *Journal of Economic Behavior* 1, 39-60.

Tversky, Amos, Paul Slovic, and Daniel Kahneman (1990). "The causes of preference reversals." *American Economic Review* 80, 204-217.

Wu, George and Richard Gonzalez (1996). "Curvature of the Probability Weighting Function." *Management Science* 42, 1676-1690.

Wu, George and Cade Massey (2003). "Learning to Detect Change: The Role of Feedback and Experience." Manuscript in preparation.