FIRST IMPRESSIONS MATTER: A MODEL OF CONFIRMATORY BIAS*

MATTHEW RABIN AND JOEL L. SCHRAG

Psychological research indicates that people have a cognitive bias that leads them to misinterpret new information as supporting previously held hypotheses. We show in a simple model that such *confirmatory bias* induces overconfidence: given any probabilistic assessment by an agent that one of two hypotheses is true, the appropriate beliefs would deem it less likely to be true. Indeed, the hypothesis that the agent believes in may be more likely to be *wrong* than right. We also show that the agent may come to believe with near certainty in a false hypothesis despite receiving an infinite amount of information.

> The human understanding when it has once adopted an opinion draws all things else to support and agree with it. And though there be a greater number and weight of instances to be found on the other side, yet these it either neglects and despises, or else by some distinction sets aside and rejects, in order that by this great and pernicious predetermination the authority of its former conclusion may remain inviolate.
>
> Francis Bacon[1]

I. INTRODUCTION

How do people form beliefs in situations of uncertainty? Economists have traditionally assumed that people begin with subjective beliefs over the different possible states of the world and use Bayes' Rule to update those beliefs. This elegant and powerful model of economic agents as Bayesian statisticians is the foundation of modern information economics.

Yet a large and growing body of psychological research

* We thank Jimmy Chan, Erik Eyster, Bruce Hsu, Clara Wang, and especially Steven Blatt for research assistance. We thank Linda Babcock, Steven Blatt, Jon Elster, Jeffrey Ely, Roger Lagunoff, George Loewenstein, and seminar participants at the University of California at Berkeley, Carnegie-Mellon University, Cornell University, Emory University, the University of Michigan, Northwestern University, the 1997 meetings of the Econometrics Society, and the 1997 meetings of the European Economic Association, as well as three referees, for helpful comments. For financial support, Rabin thanks the Alfred P. Sloan and Russell Sage Foundations, and Schrag thanks the University Research Committee of Emory University. This draft was completed while Rabin was a Fellow at the Center for Advanced Studies in the Behavioral Sciences, supported by NSF Grant #SBR-960123.

1. From *The New Organon and Related Writings* [1960; 1620], quoted in Nisbett and Ross [1980, p. 167].

The Quarterly Journal of Economics, February 1999

suggests that the way people process information often departs systematically from Bayesian updating. In this paper we formally model and explore the consequences of one particular departure from Bayesian rationality: *confirmatory bias*. A person suffers from confirmatory bias if he tends to misinterpret ambiguous evidence as confirming his current hypotheses about the world. Teachers misread performance of pupils as supporting their initial impressions of those pupils; many people misread their observations of individual behavior as supporting their prior stereotypes about groups to which these individuals belong; scientists biasedly interpret data as supporting their hypotheses.

Our simple model by and large confirms an intuition common in the psychology literature: confirmatory bias leads to overconfidence, in the sense that people on average believe more strongly than they should in their favored hypotheses. The model also yields surprising further results. An agent who suffers from confirmatory bias may come to believe in a hypothesis that is *probably wrong,* meaning that a Bayesian observer who was aware of the agent's confirmatory bias would, after observing the agent's beliefs, favor a different hypothesis than the agent. We also show that even an infinite amount of information does not necessarily overcome the effects of confirmatory bias: over time an agent may with positive probability come to believe with near certainty in the wrong hypothesis.

In Section II—which readers impatient for math may wish to skip—we review some of the psychological evidence that humans are prone to confirmatory bias. In Section III we present our formal model and provide examples and general propositions illustrating the implications of confirmatory bias. In our model, an agent initially believes that each of two possible states of the world is equally likely. The agent then receives a series of independent and identically distributed signals that are correlated with the true state. To model confirmatory bias, we assume that when the agent gets a signal that is counter to the hypothesis he currently believes is more likely, there is a positive probability that he misreads that signal as *supporting* his current hypothesis. The agent is unaware that he is misreading evidence in this way and engages in Bayesian updating that would be fully rational given his environment if he were not misreading evidence.[2]

2. Researchers have, of course, documented many other biases in information processing. We develop a model ignoring these other biases, assuming complete rationality except for this one bias, so as to keep our model tractable and because we feel that incorporating documented biases into the Bayesian model one at a time is useful for carefully identifying the effects of each particular bias.

Because we assume that the agent always correctly interprets evidence that confirms his current beliefs, relative to *proper* Bayesian updating he is biased toward confirming his current hypothesis.

So, for example, a teacher may believe either that Marta is smarter than Bart or that Bart is smarter than Marta; he initially believes each is equally likely, and over time he collects a series of signals that help him to identify who is smarter. If, after receiving one or more signals, the teacher believes that Marta is probably smarter than Bart, confirmatory bias may lead him to erroneously interpret his next signal as supporting this hypothesis. Therefore, the teacher's updated belief that Marta is smarter than Bart may be stronger than is warranted.

The notion that the teacher is likely to believe "too strongly" that Marta is smarter corresponds to the commonly held intuition that confirmatory bias leads to overconfidence. While qualifying this intuition with several caveats, our model by and large confirms it: given any probabilistic assessment by an agent that one of the hypotheses is probably true, the appropriate beliefs should on average deem it less likely to be true. Intuitively, a person who believes strongly in a hypothesis is likely to have misinterpreted some signals that conflict with what he believes, and hence is likely to have received more evidence against his believed hypothesis than he realizes.

Our analysis shows that a more surprising result arises when confirmatory bias is severe: a Bayesian observer with no direct information of her own, but who can observe the agent's belief in favor of one hypothesis, may herself believe that the *other* hypothesis is more likely. We show that such "wrongness" can arise when the agent's evidence is sufficiently mixed. Intuitively, if the agent has perceived almost as much evidence against his hypothesis as supporting it, then, since some of the evidence he perceives as supportive is actually *not* supportive, it is likely that a majority of the real signals oppose his hypothesis. Because such wrongness only arises when the agent has relatively weak evidence supporting his favored hypothesis, however, the agent *on average* correctly judges which of the two hypotheses is more likely, in the sense that his best guess is right most of the time.

While seemingly straightforward, the intuition for our overconfidence and wrongness results conceals some subtle implications of the agent's confirmatory bias. For example, an agent who currently believes in Hypothesis A (say) may *once* have believed in Hypothesis B, at which time he had a propensity to misread as

supporting Hypothesis B evidence actually in favor of Hypothesis A. But then the agent may underestimate how many signals supporting Hypothesis A he has received, and thus he may be *under*confident in his belief in favor of Hypothesis A. Indeed, we show that an agent who has only recently come to believe in a hypothesis is likely to be underconfident in that hypothesis, because until recently he has been biased against his current hypothesis. If a teacher used to think Bart was smarter than Marta and only recently concluded that Marta is smarter, then probably he has been ignoring evidence all along that Marta is smarter. The simple overconfidence and wrongness results hold because an agent has probably believed in his currently held hypothesis during most of the time he has been receiving information and so, *on average,* has been biased toward this hypothesis.

In Section IV we investigate the implications of confirmatory bias after the agent receives an infinite sequence of signals. In the absence of confirmatory bias, an agent will always come to believe with near certainty in the correct hypothesis if he receives an infinite sequence of signals. If the confirmatory bias is sufficiently severe or the strength of individual signals is weak, however, then with positive probability the agent may come to believe with near certainty that the *incorrect* hypothesis is true. Intuitively, once the agent comes to believe in an incorrect hypothesis, the confirmatory bias inhibits his ability to overturn his erroneous beliefs. If the bias is strong enough, the expected drift once the agent comes to believe in the false hypothesis is toward believing more strongly in that hypothesis, guaranteeing a positive probability that the agent ends up forever believing very strongly in the false hypothesis. The results of Section IV belie the common intuition that learning will *eventually* correct cognitive biases. While this is true for sufficiently mild confirmatory bias, when the bias is sufficiently severe "learning" can *exacerbate* the bias.

The premise of this paper is that explicit formalizations of departures from Bayesian information processing are crucial to incorporating psychological biases into economic analysis. For the most part, we do not in this paper take the important next step of developing extended economic applications of the bias we model. In Section V, however, we illustrate one implication of confirmatory bias by sketching a simple principal-agent model. We illustrate how a principal may wish to mute the incentives that she offers an agent who suffers from confirmatory bias. Indeed, we

show that even if it is very easy for an agent to gather information, so that a principal can at negligible costs provide incentives for an agent to search for profitable investment opportunities, the principal may choose *not* to provide these incentives to a confirmatory agent. This arises when the expected costs in terms of an overconfident agent investing too much in risky projects outweigh the expected benefit of the agent being better informed. We conclude in Section VI by discussing some other potential economic implications of confirmatory bias, as well as highlighting some likely obstacles to applying our model.

II. A Review of the Psychology Literature

Many different strands of psychological research yield evidence on phenomena that we are modeling under the rubric of confirmatory bias. Before reviewing this literature, we first wish to distinguish a form of "quasi-Bayesian" information processing from the bias we are examining. Although the two phenomena are related—and not always distinguished clearly in the psychology literature—they differ importantly in their implications for decision theory. Suppose that, once they form a strong hypothesis, people simply stop being attentive to relevant new information that contradicts or supports their hypotheses. Intuitively, when you become convinced that one investment strategy is more lucrative than another, you may simply stop paying attention to even freely available additional information.[3]

Bruner and Potter [1964] elegantly demonstrate such anchoring. About 90 subjects were shown blurred pictures that were gradually brought into sharper focus. Different subjects began viewing the pictures at different points in the focusing process, but the pace of the focusing process and final degree of focus were identical for all subjects. Strikingly, of those subjects who began their viewing at a severe-blur stage, less than a quarter eventually identified the pictures correctly, whereas over half of those who began viewing at a light-blur stage were able to correctly identify the pictures. Bruner and Potter [p. 424] conclude that "Interference may be accounted for partly by the difficulty of rejecting incorrect hypotheses based on substandard cues." That

3. Such behavior corresponds to a natural economic "cognitive-search" model: if we posit a cost to information processing, in many settings the natural stopping rule would be to process information until beliefs are sufficiently strong in one direction or another, and then stop.

is, people who use weak evidence to form initial hypotheses have difficulty correctly interpreting subsequent, better information that contradicts those initial hypotheses.[4]

This form of anchoring does not necessarily imply that people *misinterpret* additional evidence to either disconfirm *or* confirm initial hypotheses, only that they ignore additional evidence. Such a tendency to anchor on initial hypotheses can therefore be reconciled with Bayesian information processing. While such anchoring is potentially quite important, psychological evidence reveals a stronger and more provocative phenomenon: people tend to *misread* evidence as additional support for initial hypotheses. If a teacher initially believes that one student is smarter than another, she has the propensity to confirm that hypothesis when interpreting later performance.[5] Lord, Ross, and Lepper [1979, p. 2099] posited some of the underlying cognitive mechanisms involved in such propensities:

> . . . there is considerable evidence that people tend to interpret subsequent evidence so as to maintain their initial beliefs. The biased assimilation processes underlying this effect may include a propensity to remember the strengths of confirming evidence but the weaknesses of disconfirming evidence, to judge confirming evidence as relevant and reliable but disconfirming evidence as irrelevant and unreliable, and to accept confirming evidence at face value while scrutinizing disconfirming evidence hypercritically. With confirming evidence, we suspect that both lay and professional scientists rapidly reduce the complexity of the information and remember only a few well-chosen supportive impressions. With disconfirming evidence, they continue to reflect upon any information that suggests less damaging "alternative interpretations." Indeed, they may even come to regard the ambiguities and conceptual flaws in the data *opposing* their hypotheses as somehow suggestive of the fundamental *correctness* of those hypotheses. Thus, com-

4. A similar experiment [Wyatt and Campbell 1951] was cited by Perkins [1981] as one interpretation of the perspective that "fresh" thinkers may be better at seeing solutions to problems than people who have meditated at length on the problems, because the fresh thinkers are not overwhelmed by the "interference" of old hypotheses.

5. A related arena where the confirmation bias has been studied widely is in counselor judgments: counselors in clinical settings tend to confirm original suppositions in their eventual judgments. If you are told ahead of time that an interviewee is combative, then both your conduct and your interpretation of his conduct during an interview may reinforce that supposition, even if he is in fact no more combative than the average person. See, e.g., Haverkamp [1993]. There has also been extensive research on confirmatory bias in the interviewing process more generally; see, e.g., Dougherty, Turban, and Callender [1994] and Macan and Dipboye [1994]. Research applying variants of confirmatory bias to other domains includes Arkes [1989] and Borum, Otto, and Golding [1993] to the law; Baumann, Deber, and Thompson [1991] to medicine; and Souter [1993] discusses the implications of overconfidence to business insurance.

> pletely inconsistent or even *random* data—when "processed" in a suitably biased fashion—can maintain or even reinforce one's preconceptions.

The most striking evidence for the confirmatory bias is a series of experiments demonstrating how providing the *same* ambiguous information to people who differ in their initial beliefs on some topic can move their beliefs *farther apart*. To illustrate such polarization, Lord, Ross, and Lepper [1979] asked 151 undergraduates to complete a questionnaire that included three questions on capital punishment. Later, 48 of these students were recruited to participate in another experiment. Twenty-four of them were selected because their answers to the earlier questionnaire indicated that they were " 'proponents' who favored capital punishment, believed it to have a deterrent effect, and thought most of the relevant research supported their own beliefs. Twenty-four were opponents who opposed capital punishment, doubted its deterrent effect and thought that the relevant research supported *their* views." These subjects were then asked to judge the merits of randomly selected studies on the deterrent efficacy of the death penalty, and to state whether a given study (along with criticisms of that study) provided evidence for or against the deterrence hypothesis. Subjects were then asked to *rate,* on 16 point scales ranging from -8 to $+8$, how the studies they had read moved their attitudes toward the death penalty, and how they had changed their beliefs regarding its deterrent efficacy. Lord, Ross, and Lepper [pp. 2102–2104] summarize the basic results (all of which hold with confidence $p < .01$) as follows:

> The relevant data provide strong support for the polarization hypothesis. Asked for their final attitudes relative to the experiment's start, proponents reported that they were *more* in favor of capital punishment, whereas opponents reported that they were *less* in favor of capital punishment. . . . Similar results characterized subjects' beliefs about deterrent efficacy. Proponents reported greater belief in the deterrent effect of capital punishment, whereas opponents reported less belief in this deterrent effect.

Plous [1991] replicates the Lord-Ross-Lepper results in the context of judgments about the safety of nuclear technology. Pro- and antinuclear subjects were given identical information and arguments regarding the Three Mile Island nuclear disaster and a case of false military alert that could have led to the launching of U. S. nuclear missiles. Plous [p. 1068] found that 54 percent of pronuclear subjects became more pronuclear from the information, while only 7 percent became less pronuclear. By contrast,

only 7 percent of the antinuclear subjects became less antinuclear from the information while 45 percent became more antinuclear.[6]

Darley and Gross [1983] demonstrate a related and similarly striking form of polarization due to confirmatory bias. Seventy undergraduates were asked to assess a nine-year-old girl's academic skills in several different academic areas. Before completing this task, the students received information about the girl and her family and viewed a video tape of the girl playing in a playground. One group of subjects was given a fact sheet that described the girl's parents as college graduates who held white-collar jobs; these students viewed a video of the girl playing in what appeared to be a well-to-do, middle class neighborhood. The other group of subjects was given a fact sheet that described the girl's parents as high school graduates who held blue-collar jobs; these students viewed a video of the same girl playing in what appeared to be an impoverished inner-city neighborhood. Half of each group of subjects were then asked to evaluate the girl's reading level, measured in terms of equivalent grade level.[7] There was a small difference in the two groups' estimates—those subjects who had viewed the "inner-city" video rated the girl's skill level at an average of 3.90 (i.e., $\frac{9}{10}$ through third grade) while those who had viewed the "suburban video" rated the girl's skill level at an average of 4.29. The remaining subjects in each group were shown a second video of the girl answering (with mixed success) a series of questions. Afterwards, they were asked to

6. These percentages were derived from Table 2 of Plous [1991, p. 1068], aggregating across two studies; the remaining subjects in each case reported no change in beliefs. For other papers following on Lord, Ross, and Lepper [1979], see Fleming and Arrowood [1979]; Jennings, Lepper, and Ross [1981]; Hubbard [1984]; Lepper, Ross, and Lau [1986]. See also Miller, McHoskey, Bane, and Dowd [1993] for more mixed evidence regarding the Lord-Ross-Lepper experiment. In the passage above, Lord, Ross, and Lepper posit that even professional scientists are susceptible to such same-evidence polarization. Indeed, many economists and other academics have probably observed how differing schools of thought interpret ambiguous evidence differently. An example was once told to one of us by a colleague. He saw the same model—calibrating the elasticity of demand facing a Cournot oligopolist as a function of the number of firms in an industry—described at the University of Chicago and at the Massachusetts Institute of Technology. A Chicago economist derived the formula and said, "Look at how few firms you need to get close to infinite elasticities and perfect competition." An M.I.T. economist derived the same formula and said, "Look at how large n [the number of firms] has to be before you get anywhere close to an infinite elasticity and perfect competition." These different schools each interpreted the same *mathematical formula* as evidence reinforcing their respective views. For related analysis in the scientific domain, see also Mahoney [1977].

7. The subjects were also asked to evaluate the girl's mathematics and liberal arts skill levels; we report the results that are least supportive of the existence of confirmatory bias.

evaluate the girl's reading level. The inner-city video group rated the girl's skill level at an average of 3.71, significantly *below* the 3.90 estimate of the inner-city subjects who did not view the question-answer video. Meanwhile, the suburban video group rated the girl's skill level at an average of 4.67, significantly *above* the 4.29 estimate of the suburban subjects who did not view the second video. Even though the two groups viewed the *identical* question-and-answer video, the additional information further polarized their assessments of the girl's skill level. Darley and Gross interpret this result as evidence of confirmatory bias—subjects were influenced by the girl's background in their initial judgments, but their beliefs were evidently influenced even more strongly by the effect their initial hypotheses had on their interpretation of further evidence.[8]

Our reading of the psychology literature leads us to conclude that any of three different information-processing problems contribute to confirmatory bias. First, researchers widely recognize that confirmatory bias and overconfidence arise when people must interpret *ambiguous* evidence (see, e.g., Keren [1987] and Griffin and Tversky [1992]). Lord, Ross, and Lepper's [1979] study, discussed above, clearly illustrates the point. Keren [1988] notes the lack of confirmatory bias in visual perceptions and concludes that confirmatory tendency depends on some degree of abstraction and "discrimination" (i.e., the need for interpretation) not present in simple visual tasks. A primary mechanism of stereotype-maintenance is our tendency to interpret ambiguous behavior according to previous stereotype.[9] Similarly, a teacher may interpret an ambiguous answer by a student as either creative or just plain stupid, according to his earlier impressions of the student,

8. It should be noted that polarization of the form identified by Darley and Gross [1983] provides more direct evidence of confirmatory bias than does polarization identified by Lord, Ross, and Lepper [1979] and related papers. As Jeff Ely pointed out to us, Lord, Ross, and Lepper permit an alternative interpretation: that some people are predisposed to interpret ambiguous evidence one way and some the other. Hence, observing further polarization by groups who already differ may not reflect confirmatory bias per se, but underlying differences in interpretation of evidence that would appear irrespective of subjects' current beliefs. While this interpretation also departs from common-priors Bayesian information processing and will often yield similar implications as confirmatory bias, it is conceptually distinct and would sometimes yield different predictions. By demonstrating polarization based on differing beliefs induced in two ex ante identical groups of subjects, Darley and Gross are not subject to this alternative interpretation.

9. A vast literature explores the mechanisms by which people retain ethnic, gender, and other group stereotypes. See, e.g., Hamilton and Rose [1980]; Bodenhausen and Wyer [1985]; Bodenhausen and Lichtenstein [1987]; Stangor [1988]; Stangor and Ruble [1989]; and Hamilton, Sherman, and Ruvolo [1990].

but will be less likely to biasedly interpret more objective feedback such as answers to multiple-choice questions.

Second, confirmatory bias can arise when people must interpret statistical evidence to assess the correlation between phenomena that are separated by time. Nisbett and Ross [1980] argue that the inability to accurately identify such correlation (e.g., between hyperactivity and sugar intake, or between performance on exams and the time of day the exams are held) is one of the most robust shortcomings in human reasoning.[10] People often imagine a correlation between events when no such correlation exists.[11] Jennings, Amadibile, and Ross [1982] argue that illusory correlation can play an important role in the confirmation of false hypotheses, finding that people underestimate correlation when they have no theory of the correlation, but exaggerate correlation and see it where it is not when they have a preconceived theory of it.[12]

Third, confirmatory bias occurs when people selectively collect or scrutinize evidence. One form of "scrutiny-based" confirmatory bias is what we shall call *hypothesis-based filtering*.[13] While it is sensible to interpret ambiguous data according to current hypotheses, people tend to use the consequent "filtered" evidence

10. As Jennings, Amabile, and Ross [1982, p. 212] put it, "even the staunchest defenders of the layperson's capacities as an intuitive scientist . . . have had little that was flattering to say about the layperson's handling of bivariate observation."

11. Chapman and Chapman [1967, 1969, 1971] demonstrate that clinicians and laypeople often perceive entirely illusory correlation among (for instance) pictures and the personality traits of the people who drew the pictures. Stangor [1988] and Hamilton and Rose [1980] also discuss the role of illusory correlation in the context of confirmatory-like phenomena.

12. Similarly, Redelmeier and Tversky [1996] argue illusory correlation may help explain the persistent belief that arthritis pain is related to the weather.

13. Another mechanism can be defined as "positive test strategy": People tend to ask questions (of others, of themselves, or of data) that are likely to be true if their hypothesis is true—without due regard to the fact that they are likely to be true even if the hypothesis is *false*. See Einhorn and Hogarth [1978]; Klayman and Ha [1987]; Beattie and Baron [1988]; Devine, Hirt, and Gehrke [1990]; Hodgins and Zuckerman [1993]; Friedrich [1993]; and Zuckerman, Knee, Hodgins, and Miyake [1995]. We are using this term a bit differently than we suspect psychologists would use it. As far as we know, the term was coined by Klayman and Ha to point out that much of what was put under the rubric of confirmatory bias could indeed be a rational form of hypothesis testing. Fischhoff and Beyth-Marom [1983, pp. 255–256] and Friedrich also point out that if people are fully aware that asking "soft" questions teaches them little about the truth of hypotheses, then no bias has occurred. While we feel research on the positive test strategy needs more careful calibration versus Bayesian updating, we believe that the evidence suggests that people do not fully appreciate how little they have learned about the validity of their hypotheses when asking soft questions. (Mehle, Gettys, Manning, Baca, and Fisher [1981], for instance, show that people with specified hypotheses for observed data tend to overuse such hypotheses to explain the data because they do not have "available" the many unspecified hypothesis that could also explain the data.)

inappropriately as further evidence for these hypotheses. If a student gives an unclear answer to an exam question, it is reasonable for a teacher to be influenced in his evaluation of the answer by his prior perceptions of that student's mastery of the material. However, after assigning differential grades to students according to differential interpretation of comparable answers, it is a mistake to *then* use differential grades on the exam as *further* evidence of the differences in the students' abilities.[14] This sort of error is especially likely when the complexity and ambiguity of evidence requires the use of prior theories when interpreting data and deciding what data to examine.[15]

Finally, one of the main results in our model is confirmation of the conjecture common in the psychological literature that confirmatory bias leads to overconfidence. A vast body of psychological research, separate from research on confirmatory bias, finds that people are prone toward overconfidence in their judgments.[16]

14. Lord, Ross, and Lepper [1979, pp. 2106–2107] note a similar distinction in reflecting on the bias in their experiment discussed above. They note that it is proper for people to differentially assess probative value of different studies according to their current beliefs about the merits of the death penalty. The "sin" is in using their hypothesis-based interpretations of the strength of different studies as further support for their beliefs.

15. We suspect that hypothesis-based filtering is especially important in understanding persistence and strengthening of beliefs in tenuous "scientific" theories. Indeed, Jon Elster drew our attention to an illustration by philosopher of science Karl Popper [1963, pp. 34–35] of confirmatory bias in intellectual pursuits. Popper observed that followers of Marx, Freud, and Adler found "confirmation" everywhere, and described the process by which they strengthened their conviction over time in terms remarkably similar to the process as we've described it based on psychological research:

> Once your eyes were thus opened you saw confirming instances everywhere: the world was full of *verifications* of the theory. Whatever happened always confirmed it . . . The most characteristic element in this situation seemed to me the incessant stream of confirmations . . . As for Adler, I was much impressed by a personal experience. Once, in 1919, I reported to him a case which to me did not seem particularly Adlerian, but which he found no difficulty in analysing in terms of his theory of inferiority feelings, although he had not even seen the child. Slightly shocked, I asked him how he could be so sure. "Because of my thousandfold experience," he replied; whereupon I could not help saying: "And with this new case, I suppose, your experience has become thousand-and-one-fold."
>
> What I had in mind was that his previous observations may not have been much sounder than this new one; that each in its turn had been interpreted in the light of "previous experience," and at the same time counted as additional confirmation.

16. See, e.g., Oskamp [1982], Mahajan [1992], and Paese and Kinnaly [1993]. An early paper that makes this point is Fischhoff, Slovic, and Lichtenstein [1977], who also tested the robustness of overconfidence with monetary stakes rather than reported judgments. No decrease in overconfidence was found relative to the no-money-stakes condition. (As Camerer [1995] notes, there exist *very* few conclusions reached by researchers on judgment that have been overturned when

III. Confirmatory Bias and Belief Formation

Consider two states of the world, $x \in \{A,B\}$, where A and B are two exhaustive and mutually exclusive hypotheses regarding some issue. We consider an agent whose prior belief about x is given by prob $(x = A)$ = prob $(x = B) = 0.5$, so the agent initially views the two alternative hypotheses as equally likely to be true. In every period $t \in \{1,2,3, \ldots\}$ the agent receives a signal, $s_t \in \{a,b\}$, that is correlated with the true state of the world. Signals received at different times t are independently and identically distributed, with prob $(s_t = a|A)$ = prob $(s_t = b|B) = \theta$, for some $\theta \in (.5,1)$. After receiving each signal, the agent updates his belief about the relative likelihood of $x = A$ and $x = B$.

To model confirmatory bias, we suppose that the agent may misinterpret signals that conflict with his current belief about which hypothesis is more likely. Suppose that, given the signals the agent thinks he has observed in the first $t - 1$ periods, he believes that state A is more likely than state B. Because of his confirmatory bias, the agent may misread a conflicting signal $s_t = b$ in the next period, believing instead that he observes $s_t = a$.

Formally, in every period $t \in \{1,2,3, \ldots\}$ the agent *perceives* a signal $\sigma_t \in \{\alpha,\beta\}$. When the agent perceives a signal $\sigma_t = \alpha$, he believes that he actually received a signal $s_t = a$, and if he perceives $\sigma_t = \beta$, he believes that he actually received a signal $s_t = b$. He updates his beliefs using Bayes' Rule given his (possibly erroneous) perceptions of the signals he is receiving. We assume that with probability $q > 0$ the agent misreads a signal s_t that conflicts with his belief about which hypothesis is more likely, and that the agent always correctly interprets signals that confirm his belief. If he currently believes that Hypothesis A is more likely, then for sure he interprets a signal $s_t = a$ as $\sigma_t = \alpha$, but with probability q he misreads $s_t = b$ as $\sigma_t = \alpha$.

This model of confirmatory bias incorporates several unrealistic simplifying assumptions. For instance, we assume that the severity of the bias summarized by q does not depend on the strength of the agent's beliefs about which of the two states is more likely. It would be reasonable to expect that q is greater if the

monetary stakes are added.) There have, however, been criticisms of the evidence in support of overconfidence. See Bjorkman [1994]; Pfeifer [1994]; Tomassini, Solomon, Romney, and Krogstad [1982]; Van Lenthe [1993]; and Winman and Juslin [1993]. We feel, nevertheless, that the evidence makes a strong case for overconfidence. Indeed, see Soll [1996] for evidence that overconfidence *does* extend to ecologically valid domains.

agent's beliefs are more extreme. We conjecture that our qualitative results would continue to hold if we were to relax this assumption. Also, we assume that the agent misreads conflicting evidence as confirming evidence. While we feel that this is often the case, a reasonable alternative model would be to assume instead that the agent merely has a tendency to overlook evidence that conflicts with his beliefs. This model, too, would yield the same qualitative results as our model; intuitively, ignoring the counterhypothesis evidence in a cluster of mixed, but mostly counterhypothesis evidence, is equivalent to misreading the whole cluster as hypothesis-supportive.

The presence of confirmatory bias means that the agent's perceived signals σ_t are neither independently nor identically distributed. Suppose that, after receiving signals $s^{t-1} = (s_1, \ldots, s_{t-1})$ the agent has perceived a sequence of signals $\sigma^{t-1} = (\sigma_1, \ldots, \sigma_{t-1})$ and holds beliefs prob $(x = A|\sigma^{t-1})$. Define

$$\theta^* \equiv \text{prob}\,(\sigma_t = \alpha|\text{prob}\,(x = A|\sigma^{t-1}) > 0.5, x = B)$$
$$= \text{prob}\,(\sigma_t = \beta|\text{prob}\,(x = B|\sigma^{t-1}) > 0.5, x = A).$$
$$\theta^{**} \equiv \text{prob}\,(\sigma_t = \alpha|\text{prob}\,(x = A|\sigma^{t-1}) > 0.5, x = A)$$
$$= \text{prob}\,(\sigma_t = \beta|\text{prob}\,(x = B|\sigma^{t-1}) > 0.5, x = B).$$

θ^* and θ^{**} summarize the distribution of the agent's perceived signal σ_t when the agent believes that one hypothesis is more likely than the other; i.e., when prob $(x = A|\sigma^{t-1}) \neq 0.5$. θ^* is the probability that the agent perceives a signal confirming his belief that one hypothesis is more likely when in fact the other hypothesis is true. θ^{**} is the probability that the agent perceives a signal confirming his belief that a hypothesis is more likely when in fact it is true. Because with probability q the agent misreads a signal that conflicts with his beliefs, $\theta^* = (1 - \theta) + q\theta$ and $\theta^{**} = \theta + q(1 - \theta)$. When prob $(x = A|\sigma^{t-1}) = 0.5$, i.e., when the agent believes that the two possible hypotheses are equally likely, the agent does not suffer from confirmatory bias. In this case, he correctly perceives the signal that he receives, and he updates accurately, so $\theta \equiv$ prob $(\sigma_t = \alpha|\text{prob}\,(x = A|\sigma^{t-1}) = 0.5, x = A) =$ prob $(\sigma_t = \beta|\text{prob}\,(x = B|\sigma^{t-1}) = 0.5, x = B)$.

If $q = 0$, then the agent is an unbiased Bayesian statistician; while if $q = 1$, the agent's first piece of information completely determines his final belief, since he always misreads signals that

conflict with the first signal he receives. More generally, the higher is q, the more extreme is the confirmatory bias.

Suppose that the agent has perceived n_α α signals and n_β β signals, where $n_\alpha > n_\beta$. Because the agent believes he has received n_α a signals and n_β b signals, his updated posterior beliefs are given by

$$\text{prob}\,(x = A|n_\alpha,n_\beta) = \frac{\theta^{n_\alpha - n_\beta}}{\theta^{n_\alpha - n_\beta} + (1-\theta)^{n_\alpha - n_\beta}}.$$

Define

$$\Lambda(n_\alpha,n_\beta) = \frac{\text{prob}\,(x = A|n_\alpha,n_\beta)}{\text{prob}\,(x = B|n_\alpha,n_\beta)}.$$

$\Lambda(n_\alpha,n_\beta)$ represents the agent's beliefs in terms of a relative likelihood ratio. Using Bayes' Rule, $\Lambda(n_\alpha,n_\beta) = (\theta^{n_\alpha - n_\beta})/(1-\theta)^{n_\alpha - n_\beta}$. If $\Lambda(n_\alpha,n_\beta) > 1$, the agent believes that A is more likely than B to be the true state; while if $\Lambda(n_\alpha,n_\beta) < 1$, the agent believes that B is more likely than A. If $\Lambda(n_\alpha,n_\beta) = 1$, the agent believes that the two states are equally likely. The agent's interpretation of an additional signal is biased whenever $\Lambda(n_\alpha,n_\beta) \neq 1$.

In order to identify the effects of confirmatory bias, it is helpful to compare the agent's beliefs with the beliefs of a hypothetical unbiased, Bayesian observer who learns how many α and β signals the agent has perceived, and who knows that the agent suffers from confirmatory bias. Like the agent, the Bayesian observer initially believes that prob $(x = A)$ = prob $(x = B) = 0.5$, and she has no independent information about whether $x = A$ or $x = B$. This hypothetical observer's beliefs, therefore, reflect the true probability that $x = A$ and $x = B$, given the signals that the agent has perceived.

Define $\Lambda^*(n_\alpha,n_\beta)$ as the Bayesian observer's likelihood ratio of A versus B when she knows that an agent who suffers from confirmation bias has perceived n_α α signals and n_β β signals, where $n_\alpha > n_\beta$. In general, when $q > 0$, the biased agent's likelihood ratio $\Lambda(n_\alpha,n_\beta)$ and the unbiased observer's likelihood ratio $\Lambda^*(n_\alpha,n_\beta)$ are not equal. If $\Lambda(n_\alpha,n_\beta)$ when $n_\alpha > n_\beta$, the agent is *overconfident;* his belief in favor of the hypothesis that $x = A$ is stronger than is justified by the available evidence. Similarly, if $\Lambda(n_\alpha,n_\beta) < \Lambda^*(n_\alpha,n_\beta)$, the agent is *underconfident* in his belief that $x = A$.

In the formal results that we develop below, we assume that, while the unbiased observer knows how many α and β signals the agent has perceived, she does not know the order in which the agent perceived his signals. But when $q > 0$, the order of the agent's perceived signals, if known, would influence a Bayesian observer's beliefs, since the agent's confirmatory bias implies that his perceived signals are not distributed independently. Suppose that the agent has perceived three α signals and two β signals, in which case his beliefs are $\Lambda(n_\alpha = 3, n_\beta = 2) = \theta/(1 - \theta)$. If the Bayesian observer knew the order of the agent's signals, her posterior belief $\Lambda^*(n_\alpha = 3, n_\beta = 2)$ could be less than, greater than, or equal to $\theta/(1 - \theta)$, depending on the order of the signals. Thus, from the perspective of an outside observer, the agent could be overconfident, underconfident, or perfectly calibrated in his beliefs.

Suppose, for example, that the Bayesian observer knew that the agent's sequence of perceived signals was $(\alpha,\alpha,\alpha,\beta,\beta)$. In this case the observer's posterior likelihood ratio is

$$\Lambda^* = \frac{\theta(\theta + q(1 - \theta))^2(1 - \theta)^2}{(1 - \theta)(1 - \theta + q\theta)^2\theta^2}$$

$$= \frac{(\theta + q(1 - \theta))^2(1 - \theta)}{(1 - \theta + q\theta)^2\theta} < \frac{\theta}{1 - \theta}, \quad \forall q \in (0,1].$$

Intuitively, the Bayesian observer recognizes the possibility that the agent may have misread his second and third signals, perceiving that they supported the hypothesis that $x = A$ when in fact one or both may have supported the hypothesis that $x = B$. Therefore, the Bayesian observer is less convinced that $x = A$ than the agent, who is overconfident in his belief. More generally, an observer who knows that a biased agent has always believed in his current hypothesis should judge the agent to be overconfident in his belief, since there is a positive probability that the agent has misread signals that are counter to his favored hypothesis. An observer who knows that a teacher has always believed that Bart is smarter than Marta should recognize that the teacher's confirmatory bias may have led him to misread evidence that Marta is in fact smarter.

Alternatively, suppose that the Bayesian observer knew that the agent's sequence of perceived signals was $(\beta,\beta,\alpha,\alpha,\alpha)$. Now the

observer's posterior likelihood ratio is

$$\Lambda^* = \frac{(1-\theta)(1-\theta+q\theta)\theta^3}{\theta(\theta+q(1-\theta))(1-\theta)^3}$$

$$= \frac{(1-\theta+q\theta)\theta^2}{(\theta+q(1-\theta))(1-\theta)^2} > \frac{\theta}{1-\theta}, \quad \forall q \in (0,1].$$

In this case, the Bayesian observer believes that the agent may have misread his second signal, perceiving that it supported the hypothesis that $x = B$ when in fact it may have supported the hypothesis that $x = A$. Thus, the Bayesian observer believes that there is a greater likelihood that $x = A$ than the agent, who is *under*confident in his belief. More generally, an observer who knows that a biased agent only recently came to believe in his current hypothesis after long believing in the opposite hypothesis should judge the agent to be underconfident in his belief, since the agent may have misread one or more signals that support his current hypothesis when he believed the opposite. An observer who knows that a teacher initially thought that Bart was smarter than Marta, but eventually started to believe that it was slightly more likely that Marta was smarter than Bart, should conjecture that the teacher is underconfident about his new hypothesis. When the teacher believed that Bart was smarter than Marta, he may have misinterpreted signals that Marta was smarter. The fact that the teacher came to believe that Marta was smarter *despite* his initial bias toward believing that Bart was smarter indicates that the evidence is very strong that Marta is smarter.[17]

The preceding examples illustrate how information about the order of the agent's signals would significantly influence an

17. As a discussant for this paper, Roger Lagunoff made an interesting suggestion that is especially relevant for the examples we are discussing here. In our model, once the agent interprets a signal, he never goes back and reinterprets it—even if he later changes his hypothesis about the world. Hence we are not capturing a form of belief updating we sometimes observe: when somebody (finally) comes around to change his world-view that he held for quite a while, he sometimes experiences an epiphany whereby he goes back and reinterprets previous evidence in light of his new hypothesis, realizing that "the signs were there all along." This suggests a model in which an agent is biased in interpreting not just the next signal, but all past signals, as supporting his current hypothesis about the world. While we suspect there is some truth to this, we don't believe that people fully retroactively rebias themselves in this way. (We have found no psychological evidence about this one way or another.) While such an alternative model would rule out the possibility of "underconfidence" for recent converts, it would leave all the predictions regarding overconfidence discussed in the remainder of the paper qualitatively the same, and magnify the magnitude of our results (and simplify the proofs).

outside observer's judgment about whether, and in what direction, the agent's beliefs were biased. Nevertheless, for the remainder of the paper we assume that an outside observer only knows the number of α and β signals that the agent has received, and not the order in which he received them. This assumption enables us to identify whether, on *average,* the agent is over- or underconfident. This appears to be the question that the psychological literature addresses; presumably, it is also of interest to economists.

Clearly, if $q = 0$, then $\Lambda^*(n_\alpha,n_\beta) = \Lambda(n_\alpha,n_\beta)$. When $q > 0$, however, Proposition 1 establishes that $\Lambda^*(n_\alpha,n_\beta) < \Lambda(n_\alpha,n_\beta)$. That is, when the agent perceives that a majority of his signals support (say) Hypothesis A, he believes in A with higher probability than is warranted.[18]

PROPOSITION 1. Suppose that $n_\alpha > n_\beta$ and $n_\alpha + n_\beta > 1$. Then $\Lambda^*(n_\alpha,n_\beta) < \Lambda(n_\alpha,n_\beta)$.

Proposition 1 establishes that an agent who suffers from confirmatory bias will be overconfident in his belief about which state is most likely.

An observer who knows the agent's beliefs cannot usually observe the exact sequence of the agent's perceived signals. Therefore, the observer's judgment about whether the agent is under- or overconfident depends on her belief regarding the likelihood of the different possible sequences of signals. Proposition 1 establishes that overconfidence is the dominant force. The intuition for this result is fairly straightforward: if you cannot directly observe the agent's past beliefs, but you know that he now believes in Hypothesis A, you should surmise that, on average, he spent more time in the past believing Hypothesis A than Hypothesis B. Consequently, you should surmise that, on average, the agent misread more signals while believing in Hypothesis A—contributing to overconfidence—than he misread while believing in Hypothesis B—contributing to underconfidence. Proposition 1 hinges to some extent on our assumption that the agent receives signals that are the same strength in every period. We believe that (far more complicated) versions of Proposition 1 hold in more general models, but we show in Appendix 1 that *under*confidence is sometimes possible when the agent's signals are of different strengths in different periods.

18. All proofs are in Appendix 2. Because our model is entirely symmetric, we shall for convenience present all results and much of our discussion solely for the case where A is perceived as more likely.

Proposition 1 shows that when the agent believes that the state is $x = A$ with probability $\mu > 0.5$, the true probability that the state is $x = A$ is less than μ. Interestingly, the true probability that A is the true state *may be less than* 0.5, meaning that B is more likely than A. The possibility that the agent may suffer not merely from overconfidence, but also from "wrongness," arises when the agent's confirmatory bias is severe and he has perceived at least two signals in favor of each hypothesis.

To see the intuition for this result, suppose that the agent has since his first signal $s_1 = a$ believed that Hypothesis A is more likely than B, but that he nevertheless has perceived two signals $\sigma_t = \sigma_t = \beta$ at two times $t, t' > 1$. If the agent's confirmatory bias is severe (i.e., $q \approx 1$), only his first perceived signal in favor of A provides true evidence that $x = A$. Once the agent believes that A is true, his confirmatory bias predisposes him to perceive that subsequent signals support this belief, and, therefore, additional signals in favor of A are not very informative. But, because the agent's two perceived signals in favor of B conflict with what he believes—that $x = A$ is more likely—they reflect *actual* signals in favor of B. Thus, although the agent has always believed that $x = A$ is more likely, he has effectively received only one signal in favor of A and two signals in favor of B. In this case the agent's belief that $x = A$ represents extreme overconfidence; if he had correctly interpreted evidence, he would believe that $x = B$ is more likely.

It is, of course, possible that hypothesis A is *more* likely than the agent realizes if he first perceives a signal $s_1 = b$, falsely reads a's as b's for a while, and only later perceives enough α's to come to believe in A. And it is true that getting more true a's than true b's implies that Hypothesis A is more likely. Yet, it can be shown that these possibilities may be far less likely than the cases leading to extreme overconfidence, so that the net effect that is more likely that B is true than that A is true if the agent believes in A with mixed evidence.

For example, suppose that the agent has perceived seven signals, four α's and three β's. Given these signals, the agent's posterior beliefs are $\Lambda(n_\alpha = 4, n_\beta = 3) = \theta/(1 - \theta) > 1$; the agent believes that the state $x = A$ is more likely. Meanwhile, the true likelihood ratio is $\Lambda^*(n_\alpha = 4, n_\beta = 3) =$

$$\frac{\begin{array}{r}(1-\theta)^3[8\theta^4 + 8\theta^3\theta^{**} + 7\theta^2\theta^{**2} + 5\theta\theta^{**3}] \\ + (1-\theta)^2[\theta^3\theta^{**}\theta^* + 4\theta^4\theta^*] + 2\theta^4(1-\theta)\theta^{*2}\end{array}}{\begin{array}{r}\theta^3[8(1-\theta)^4 + 8(1-\theta)^3\theta^* + 7(1-\theta)^2\theta^{*2} + 5(1-\theta)\theta^{*3}] \\ + \theta^2[(1-\theta)^3\theta^{**}\theta^* + 4(1-\theta)^4\theta^{**}] + 2\theta(1-\theta)^4\theta^{**2}\end{array}}.$$

Suppose that $\theta = .75$. Then the agent's posterior likelihood ratio is $\Lambda(4,3) = 3$. Suppose further that $q = .95$, and therefore the agent suffers from severe confirmatory bias. Then, the true likelihood ratio is $\Lambda^*(4,3) = .63$, and therefore $x = B$ is more likely to be the true state, despite the agent having perceived more α signals than β signals.

Indeed, it turns out to be the case that when confirmatory bias is very severe *and* the signals are very informative, then whenever you observe the agent believing in Hypothesis A and having perceived two or more β signals, then you should assume that it is more likely that B is true than A. We formalize this in Proposition 2. Let $\Lambda^*(n_\alpha,n_\beta|q,\theta)$ be the appropriate beliefs as a function of q and θ. Then

PROPOSITION 2. For $n_\alpha > n_\beta$ and $n_\beta \leq 1$, $\lim_{\epsilon \to 0} \Lambda^*(n_\alpha,n_\beta|1-\epsilon, 1-\epsilon) > 1$. For all $n_\alpha > n_\beta \geq 2$, $\lim_{\epsilon \to 0} \Lambda^*(n_\alpha,n_\beta|1-\epsilon,1-\epsilon) < 1$.

That is, for θ and q both very close to 1, when the agent has perceived one or fewer β signals and believes in Hypothesis A, she is probably correct (though overconfident) in her beliefs; when the agent has perceived two or more β signals and believes in Hypothesis A, she is probably *in*correct in her beliefs—Hypothesis B is more likely to be true.

We emphasize that the very premise of the proposition means that the situations to which it applies are uncommon; when both q and θ are close to 1, the probability of perceiving anything besides a sequence of signals favoring the correct hypothesis is small. Therefore, Proposition 2 tells us about a very low-probability event. In our example with seven signals, $q = .95$, and $\theta = .75$, the probability that the signals are sufficiently mixed that the agent is probably wrong is a little more than one-half percent.

While we do not know more generally the highest probability with which the agent can be wrong, some calibrations illustrate that it *can* be relatively likely that the agent ends up with beliefs that a Bayesian observer would deem probably wrong. Tables I–IV display, for various values of n, θ, and q, the probability that $\Lambda^* < \eta$ and $\Lambda > 1$ or $\Lambda^* > 1/\eta$ and $\Lambda < 1$, where η represents different thresholds for how wrong the agent is. Table entries are in percentage terms (rounded to the nearest percent), with rows corresponding to different values of q and columns to different values of θ. (Dashes indicate an entry *exactly* equal to zero.)[19]

For instance, with $\theta = .6$ and $q = .5$, the probability that the

19. The entries in Tables I–IV reflect direct calculations (performed by computer) of the probabilities in question.

TABLE I
PROBABILITY OF "WRONGNESS," $n = 50$, $\eta = 1$

	θ			
q	0.6	0.7	0.7	0.9
.1	—	—	—	—
.2	12	2	0	—
.3	21	9	1	0
.4	29	15	5	1
.5	27	18	10	3
.6	33	22	12	5
.7	27	21	15	7
.8	33	24	15	8
.9	21	17	12	9

TABLE II
PROBABILITY OF "WRONGNESS," $n = 50$, $\eta = ½$

	θ			
q	0.6	0.7	0.7	0.9
.1	—	—	—	—
.2	—	—	—	—
.3	10	5	1	—
.4	15	13	5	1
.5	19	18	9	3
.6	16	18	12	5
.7	18	21	13	7
.8	11	16	15	7
.9	4	8	12	6

TABLE III
PROBABILITY OF "WRONGNESS," $n = 50$, $\eta = ⅑$

	θ			
q	0.6	0.7	0.7	0.9
.1	—	—	—	—
.2	—	—	—	—
.3	—	—	—	—
.4	—	5	—	—
.5	3	12	7	1
.6	3	14	11	4
.7	2	11	12	6
.8	1	6	12	7
.9	0	4	7	6

TABLE IV
PROBABILITY OF "WRONGNESS," $n = 7$, $\eta = 1$

	θ			
q	0.6	0.7	0.7	0.9
.1	—	—	—	—
.2	—	—	—	—
.3	—	—	—	—
.4	—	—	—	—
.5	—	—	—	2
.6	—	5	3	1
.7	3	3	2	5
.8	10	8	6	3
.9	3	2	2	1

agent has beliefs after 50 signals that the observer would deem probably wrong is about 27 percent. The probability in this same case that his beliefs will lead the observer to believe in the other hypothesis with at least probability ⅔ is 19 percent, and the probability that the observer would believe in the hypothesis opposite to the agent's with at least 9/10 probability is about 3 percent.[20]

In the example above and in Proposition 2, the agent can be wrong in her beliefs. Even more surprising, perhaps, the true probability that A is the correct hypothesis need not be monotonically increasing in the proportion of α signals the agent perceives. Continue to assume that the agent has received seven signals, but now suppose that five support $x = A$ and two support $x = B$. Then, because $\theta = .75$, the agent's posterior likelihood ratio is $\Lambda(5,2) = 27 > \Lambda(4,3)$. Meanwhile, the true likelihood ratio is

$$\Lambda^*(n_\alpha = 5, n_\beta = 2) = \frac{(1-\theta)^2[7\theta^2\theta^{**3} + 9\theta\theta^{**4} + 4\theta^3\theta^{**2}] + (1-\theta)\theta^3\theta^{**2}\theta^*}{\theta^2[7(1-\theta)^2\theta^{*3} + 9(1-\theta)\theta^{*4} + 4(1-\theta)^3\theta^{*2}] + \theta(1-\theta)^3\theta^{*2}\theta^{**}}.$$

20. Readers may note that these probabilities generally increase in q and then decrease, with probability about 0 for $q = 0$ and $q = 1$. But they are not single-peaked in q. This is because there are two factors at work in determining the influence of q on the probability. As q increases, the probability that the agent will end up with close-to-even mixes of α and β signals decreases continuously. But because an increase in q increases the likelihood that any given combination of α's and β's involves the agent being probabilistically wrong, there will be at certain points discrete jumps upward in the likelihood of wrongness for some values of q. The result is an extremely poorly behaved function.

Maintaining the assumption that $q = .95$, $\Lambda^*(5,2) = .62 < \Lambda^*(4,3) = .63$. Therefore, the relative likelihood that the true state is $x = A$ versus $x = B$ is smaller if the agent perceives that five out of seven signals support $x = A$ than if he perceives that only four out of seven signals support $x = A$.

While seemingly counterintuitive, this result reflects the fact that the agent is more likely to have perceived (truly informative) signals $\sigma_j = \beta$ that conflict with a belief that $x = A$ when he has perceived only two signals in favor of B than when he has perceived three signals in favor of B. Intuitively, the agent is more likely to have believed for many periods that $x = A$ in the former case than in the latter case. Put differently, the agent is *less likely* to have perceived (truly informative) signals $\sigma_j = \alpha$ that conflict with a belief that $x = B$ when he has perceived only two signals in favor of B than when he has perceived three signals in favor of B.

The preceding examples illustrate that an agent who suffers from confirmatory bias may believe that one of the two possible states is more likely than the other when in fact the reverse is true. Nevertheless, Proposition 3 shows that a Bayesian observer who knows only that a biased agent believes that $x = A$ is more likely than $x = B$ will herself believe that $x = A$ is more likely. Therefore, an agent who suffers from confirmatory bias will "on average" correctly judge which of the two possible states is more likely, though, as Proposition 1 establishes, he will always be overconfident in his belief.

Define $\Lambda^*(n)$ as the likelihood ratio of a Bayesian observer who knows that a confirmatory agent has perceived a total of n signals, and knows that $n_\alpha > n_\beta$, but does not know the exact values of n_α and n_β. That is, the observer knows only that the agent believes A is more likely than B, but observes nothing about the strength of his beliefs. Then

PROPOSITION 3. For all n, $\Lambda^*(n) > 1$.

In light of the above examples where the agent may be wrong, the simple generality of Proposition 3 may seem surprising. It is reconciled with the examples by observing that the agent suffers from "wrongness" only when his confirmatory bias is very severe, meaning that q is close to 1, and yet he has perceived mixed signals about which state is more likely. But the agent is unlikely to receive mixed signals when his confirmatory bias is strong, because each signal σ_t will tend to mirror σ_1.

IV. Beliefs after an Infinite Number of Signals

A fully Bayesian agent—for whom $q = 0$—will after an infinite number of signals come to believe with near certainty in the correct hypothesis. We now investigate the implications of confirmatory bias in the limit as an agent receives an infinite number of signals.

We begin with definitions and a lemma that will help to analyze this question. Suppose that the agent has thus far received $m = n_\alpha - n_\beta > 0$ more perceived signals in support of Hypothesis A than in support of Hypothesis B. Suppose further that, as long as $n_\alpha > n_\beta$, prob $(\sigma_t = \alpha) = \gamma$. Note that $\gamma = \theta^*$ if B is true, and $\gamma = \theta^{**}$ if A is true. We wish to consider some preliminary results that hold in either case. We define $p(m,\gamma)$ as the probability that there exists some time in the future when the agent will have received an equal number of α and β signals. (At that time the agent's posterior belief is the same as his prior belief, prob $(x = A) = 0.5$.) We have the following lemma, which is a restatement of a well-known result from Feller [1968, pp. 344–347).

Lemma 1. For all $m > 0$, $\gamma \geq 0.5$, $p(m,\gamma) = [(1 - \gamma)/\gamma]^m$. For $\gamma \leq 0.5$, $p(m,\gamma) = 1$.

We define P_W as the probability that the agent, beginning with the prior belief prob $(x = A) = 0.5$, comes to believe with certainty in the wrong hypothesis after receiving an infinite number of signals.[21] That is, P_W is the probability that, although the true state is $x = A$, the agent instead comes to believe irreversibly, with near certainty, that $x = B$. Proposition 4 characterizes P_W as a function of q and θ.

Proposition 4. If $q > 1 - 1/(2\theta)$, then

$$P_W = \frac{(1 - \theta) \cdot (1 - (1 - \theta^*)/\theta^*)}{(1 - (1 - \theta) \cdot ((1 - \theta^*)/\theta^*) - \theta((1 - \theta^{**})/\theta^{**}))} > 0.$$

If $q \leq 1 - 1/(2\theta)$, then $P_W = 0$.

When $q > 1 - 1/(2\theta)$, $\theta^* = (1 - \theta) + q\theta > 0.5$. When $\theta^* > 0.5$, then once the agent comes to believe that the wrong hypothesis about x is more likely, he is consequently more likely to receive a

21. Formally, $P_W \equiv$ prob $(\forall k > 0$ and $\forall \epsilon > 0, \exists n^*$ such that prob $(n_\alpha - n_\beta > k$ for all $n > n^*) > 1 - \epsilon)$.

TABLE V
PROBABILITY OF BELIEVING IN WRONG HYPOTHESIS AFTER OBSERVING AN INFINITE NUMBER OF SIGNALS

q	$\theta = .6$	$\theta = .667$	$\theta = .75$	$\theta = .9$
.25	18	—	—	—
.333	26	12	—	—
.5	34	24	13	2
.75	38	31	22	8

signal σ_t that *confirms* this incorrect belief than he is to receive a signal that *conflicts* with this incorrect belief. This guarantees that there is a positive probability that the agent will never overturn his incorrect hypothesis, and in fact come to believe more and more strongly in that wrong hypothesis. Conversely, if $q < 1 - 1/(2\theta)$, then $\theta^* < .5$, which guarantees that the agent will, every time he comes to believe the wrong hypothesis is more likely, eventually come to abandon that belief. This in turn implies that the agent will repeatedly come to believe the correct hypothesis is more likely; and since $\theta^{**} = \theta + q(1 - \theta) > \theta > .5$, he will eventually come to believe in it with near certainty.

The proposition shows that, despite receiving an infinite number of signals, the agent may become certain that the incorrect hypothesis is in fact true.[22] This occurs when the agent's confirmatory bias is sufficiently severe. To illustrate the magnitude of P_W, Table V displays P_W for various values of θ and q. Table entries are in percentage terms (rounded to the nearest percent), with rows corresponding to different values of q and columns to different values of θ. (Dashes indicate an entry *exactly* equal to zero.)

For example, suppose that $q = 0.5$ and $\theta = .75$. Then $P_W = 7/52$, meaning that approximately 13 percent of the time the agent will eventually come to believe with certainty in the wrong hypothesis. As the quality of the agent's true signal worsens he is more likely to believe with certainty in the wrong hypothesis. Indeed, a corollary to Proposition 4 is that, fixing any $q > 0$, $\lim_{\theta \to 1/2} P_W = 1/2$.

We now investigate the related question of when the agent will maintain an incorrect initial belief. To do so, we relax our

22. It is straightforward to show that the agent becomes certain that the *correct* hypothesis about the state of the world is true with complementary probability. Therefore, after an infinite number of signals the agent will believe that *one* of the hypotheses is certainly true.

assumption that the agent initially believes that each state x is equally likely and suppose instead that the agent initially believes that the wrong hypothesis is more likely to be true. For example, if $x = B$ is the true state of the world, then the agent initially believes prob $(x = A) = \mu > 0.5$. Crucially, we assume that this belief arose from signals that are independent of the new signals that the agent receives, which are distributed as outlined above.

Given the assumption that the signals are independently distributed and ignoring integer problems, these prior beliefs can be interpreted as if the agent has already received D more signals supporting the incorrect hypothesis, where

$$\mu = \theta^D/[\theta^D + (1 - \theta)^D].$$

This formula implicitly defines a function $D(\mu)$. The agent must receive $D(\mu)$ more conflicting signals σ_t than confirming signals in order to reach a posterior belief that the two possible states of the world, A and B, are equally likely.

We define $P_W(\mu)$ as the probability that the agent, beginning with the prior belief $\mu > 0.5$ that the wrong hypothesis about the state of the world is true, comes to believe with certainty in the wrong hypothesis after receiving an infinite number of signals.[23]

PROPOSITION 5. Choose any $\epsilon > 0$ and any $\mu > 0.5$. Then

(i) For all $\theta \in (0.5,1)$, there exists $q > 0$ such that $PW(\mu) > 1 - \epsilon$.

(ii) For all $q > 0$, there exists $\theta > 0.5$ such that $PW(\mu) > 1 - \epsilon$.

Proposition 5 says that an agent who begins with an arbitrarily small bias in the direction of the incorrect hypothesis will almost surely maintain his belief in this hypothesis when either of two conditions is satisfied. First, and not very surprisingly, this will occur when the agent is subject to severe confirmatory bias. When q is very close to 1, then the agent almost never receives signals that conflict with his initial belief, and therefore it is not surprising that this belief is rarely overturned. Second, and somewhat more surprisingly, the agent almost surely maintains his incorrect belief provided that his true signals are very weak, meaning that θ is very close to 0.5. This result does not depend on the level of confirmatory bias, so long as $q > 0$. This result means that if the agent receives only very weak feedback from his environment and is subject to any confirmatory bias, he almost

23. $P_W(0.5) = P_W$.

never overcomes any initial beliefs that are significantly incorrect, and in fact comes to believe that the incorrect hypothesis is certainly true. While one should not overinterpret the second result in Proposition 5—we can question whether agents really pay attention to such weak feedback—the conclusion is nevertheless very striking. Propositions 4 and 5 show that an infinite sequence of signals will not necessarily lead people to overcome erroneous beliefs; rather, people may simply become more and more confident in those erroneous beliefs.

Table VI displays $P_W(\mu)$ for various values of θ, q, and μ. If θ and μ are chosen in such a way that $D(\mu)$ is an integer, and $q > 1 - 1/2\theta$, it follows from Lemma 1 that

$$P_W(\mu) = \left(1 - \left[\frac{1-\theta^*}{\theta^*}\right]^{D(\mu)}\right) + \left(P_W(0.5)\left[\frac{1-\theta^*}{\theta^*}\right]^{D(\mu)}\right) > 0.$$

Table entries are in percentage terms (rounded to the nearest percent), with rows corresponding to different values of q and columns to different values of θ and the prior belief μ. (Dashes indicate an entry *exactly* equal to zero.)

For example, suppose that $\theta = .551$ and $\mu = .6$. In this case $D(\mu) = 2$, meaning that the agent must receive two more signals that conflict with rather than confirm his prior belief in order to believe that the states A and B are equally likely. But if $q = .333$, there is nearly an 80 percent chance that the agent will never overturn his incorrect prior belief. Clearly, learning does not necessarily lead the agent to correctly identify the true state.

V. Confirmatory Bias in a Principal-Agent Model

Confirmatory bias is likely to influence economic behavior in many different arenas. In this section we develop a simple

TABLE VI
Probability of Maintaining an Incorrect Prior Belief after Receiving an Infinite Number of Signals

q	$\theta = .6, \mu = .6,$ $D(\mu) = 1$	$\theta = .551, \mu = .6,$ $D(\mu) = 2$	$\theta = .75, \mu = .75,$ $D(\mu) = 1$	$\theta = .634, \mu = .75,$ $D(\mu) = 2$
.25	32	67	—	24
.33	51	79	—	56
.5	72	92	48	85
.75	89	99	82	98

illustrative model of a principal-agent relationship, a context where we think confirmatory bias is likely to be important. The premise of the model is that an agent may take inappropriate actions not solely because of intentional misbehavior—moral hazard—but also because of unintentional errors arising from confirmatory bias. Specifically, because an agent who suffers from confirmatory bias will be overconfident in his judgment about how likely various actions are to pay off, he may be prone to taking actions that are riskier and more "extreme" than is optimal for the principal. Such overconfidence seems to reflect the intuition among some researchers: at a conference one of the authors attended, a leading economist conjectured that bad investment decisions by businesses in Eastern Europe receiving bank loans were more often the result of overconfidence by borrowers than of intentions to mislead banks. Even more directly along the lines of our model, Wood [1989] asserts that money managers become more confident in their investment decisions as they gather more information—even when the quality of their investment decisions is not improved.

A principal who is aware of an agent's confirmatory bias will wish to design incentives that both cause the agent to internalize the negative consequences of bad choices *and* prevent decisions based on good-faith overconfidence. In particular, incentives that lead the agent to collect a lot of information may not be optimal if the agent suffers from severe confirmatory bias and, hence, becomes more overconfident as he collects more information. The principal may therefore wish to mute the agent's incentives relative to what would be optimal in the absence of confirmatory bias.

While an exhaustive analysis of the effect of confirmatory bias on agency relationships is beyond the scope of this paper, we now develop a simple illustrative model along these lines. Suppose that a principal hires an agent to allocate initial wealth $W = 1$ between the different investments in the set $I = \{I_A, I_B, I_C\}$. The investment I_C is risk-free; it always yields a gross return $r(I_C) = 1$. Investments I_A and I_B, on the other hand, are risky; their returns depend on the state of nature $x \in \{A,B\}$. Conditional on the state x, the gross returns from I_A and I_B are $r(I_A|A) = r(I_B|B) = R \in (1,2)$ and $r(I_A|B) = r(I_B|A) = 0$. Let $u(\cdot)$ be the principal's Von Neumann-Morgenstern utility function for money. Because the principal may be risk-averse, we assume that $u' > 0$ and $u'' \leq 0$. Consistent with the model that we developed in the previous sections, the

principal and the agent cannot observe the state x, and they hold a common prior belief that prob $(x = A)$ = prob $(x = B)$ = 0.5. Hence, if the agent learns nothing more about the true state, the optimal investment is in the riskless investment I_C; if he learns sufficiently more—generating beliefs sufficiently different from .5—he will perceive it as optimal to invest some money in one of the two risky investments.[24]

Before choosing how to invest the principal's wealth, the agent has the opportunity to observe informative signals about the true state, although the agent's confirmatory bias may lead him to misinterpret these signals. We assume that the signals that the agent receives, and the way he perceives these signals, accord with the model in the previous sections.

For both analytic ease and to highlight the role of confirmatory bias, we abstract away from the usual moral-hazard concerns: we assume that the agent costlessly observes signals about the state x and expends no effort when making decisions on the principal's behalf. Under this assumption, an arbitrarily small incentive to identify the true state would lead the agent to observe an infinite number of signals, after which he would believe that he could identify the true state with near certainty. Furthermore, to abstract away from issues of optimal risk-sharing between the agent and the principal, we assume that the agent is (nearly) infinitely risk-averse. Therefore, the principal must offer the agent a nearly constant wage.[25] Under these assumptions, the

24. While the language and notation suggest that we are referring to well-defined investment portfolios (e.g., three different bonds), we mean for the model to apply as well to internal organizational incentives to pursue ambiguously defined projects. Indeed, this alternative interpretation may better fit the formal model in some respects. Note that it is crucial to our analysis that the agent cannot or does not merely report his beliefs to the principal, but rather implements a strategy himself based on his beliefs. If the principal knew the agent's beliefs and the extent of his confirmatory bias, she could form her own beliefs about the true state of the world and then directly choose the action that would maximize her expected payoff.

25. These assumptions raise a subtle point. If the agent anticipated gathering an infinite number of signals, he would be willing to accept a contract that yielded a payoff that depended on the outcome of a risky investment, even if he were infinitely risk-averse. This is because the agent would anticipate being able to identify the true state with virtual certainty. But, if the hypothesis of the first part of Proposition 4 is satisfied, the agent will be overconfident in his judgment after observing an infinite number of signals. Therefore, such a contract would impose more risk on the agent—and yield a lower expected utility—than he anticipated. We assume here that the principal cannot exploit the agent's confirmatory bias by convincing him to sign a contract that yields an expected payoff that is less than the agent's reservation payoff. This assumption means that the principal cannot use the agent as a "money pump," and its empirical validity deserves investigation. The issue of whether to focus analysis on "efficiency contracts" rather than "money-pumping contracts" is a more general one that is likely to arise in

incentives that the principal offers to the agent affect the principal's payoff only through their effect on the investment decision that the agent makes. The principal does not need to compensate the agent for gathering information, and she cannot transfer risk to the agent.

These assumptions permit us to focus on two polar cases. In the first case, the principal gives the agent no incentive to collect information, and the agent allocates the principal's wealth with no information beyond his prior belief. In the second case, the principal gives the agent an arbitrarily small incentive to collect information, and the agent allocates the principal's wealth after observing an infinite number of signals and coming to believe with virtual certainty that he has identified the true state.

Suppose first that the principal offers the agent no incentive to gather information. Because $R < 2$ and the principal's prior belief is prob $(x = A) = 0.5$, it is optimal for the principal to direct the agent to invest all of the principal's wealth in the risk-free asset I_C. Now suppose that the principal offers the agent a small incentive to identify the true state. For instance, suppose that the principal offers to pay the agent an arbitrarily small fraction of the principal's gross return. Because it is free for the agent to collect information, he would observe an infinite number of signals and come to believe that he could identify the true state with certainty. These extreme beliefs would lead the agent to allocate all of the principal's wealth to one of the risky assets. If, for example, the agent thought that A was surely the true state, he would allocate all of the principal's wealth to asset I_A.[26]

If $q \leq 1 - 1/(2\theta)$ (i.e., if confirmatory bias is sufficiently weak), Proposition 4 establishes that the agent will eventually identify the true state with near certainty if he collects enough signals. Hence, under our assumption that it is costless for the agent to gather information and that in every state of the world one or the other of the "risky" investments is optimal, it would then be optimal for the principal to offer a contract that would lead the agent to collect an infinite number of signals, and the principal would receive a payoff $u(R) > u(1)$.

If $q > 1 - 1/(2\theta)$, on the other hand, Proposition 4 establishes

developing formal models of incentives for boundedly rational agents. See, for instance, O'Donoghue and Rabin [1997], who study incentive design for agents who irrationally procrastinate, and who discuss various rationales for focusing on efficiency contracts.

26. We assume that short-selling is impossible, so the agent could not allocate more than \$1 to asset α by, for instance, selling investment I_C short.

that the agent's (completely confident) belief about the true state is wrong with positive probability. It can be shown that he *correctly* identifies the true state with probability

$$\mu^*(\theta,q) = \frac{\theta[2(\theta + q(1 - \theta)) - 1](1 - \theta + q\theta)}{q[1 - 2(1 - q)\theta(1 - \theta)]}.$$

Since the agent is fully confident that he has identified the true state, he invests all of the principal's wealth in the risky asset he believes is most profitable, so the principal's payoff is $\mu^*(\theta,q)u(R) + (1 - \mu^*(\theta,q))u(0)$. Note that $\mu^*(\theta,q)$ is increasing in θ and decreasing in q, meaning that the agent identifies the true state with higher probability when he receives more informative signals and lower probability when his confirmatory bias is more severe.

The principal does not want the agent to become "informed" when $u(1) \geq \mu^*(\theta,q)u(R) + (1 - \mu^*(\theta,q))u(0)$. Define $\underline{\mu}$ as satisfying $u(1) = \underline{\mu}u(R) + (1 - \underline{\mu})u(0)$; if the agent correctly identifies the true state with probability $\underline{\mu}$ after observing an infinite number of signals, the principal is just willing for the agent to become informed about the state x. Because $\mu^*(\theta,q) \geq \theta$, the principal always offers the agent an incentive to become informed if $\theta \geq \underline{\mu}$. That is, if the principal prefers all her money invested in a risky investment based on just one signal to having all her money invested in the risk-free investment, she will provide incentives to the agent. When $\theta < \underline{\mu}$, on the other hand, we have Proposition 6.

PROPOSITION 6. Suppose that $\theta < \underline{\mu}$.

(i) There exists $q^* \in (1 - 1/2\theta,1]$ such that the principal does not offer the agent an incentive to become informed about the state x if and only if $q \geq q^*$.
(ii) For any $q \in [0,1]$, there exists $\theta^* \in (0.5,\underline{\mu}]$ such that the principal does not offer the agent an incentive to become informed about the state x if and only if $\theta \leq \theta^*$.

When $\theta < \underline{\mu}$, the principal does not want the agent to observe signals about the state x either if confirmatory bias is very severe or if the agent receives very weak signals. In either case, there is a strong possibility that an "informed" agent would erroneously identify the true state, although the agent himself would overconfidently believe that he could identify the true state with near certainty. If the agent's overconfidence is sufficiently severe, the principal prefers not to offer the agent any incentive to become informed, in which case the agent will invest the principal's wealth in the riskless asset.

The principal's degree of risk aversion also influences whether or not he wants the agent to observe signals about the state x. While Proposition 6 shows that there are conditions where even a risk-neutral principal eschews incentives for the agent, the principal is more bothered by the overconfidence when she is more risk-averse, in the usual sense defined by Pratt [1964]. Indeed, whenever confirmatory bias is severe enough that the agent might be wrong even after gathering an infinite number of signals, a principal who is sufficiently risk-averse will prefer not to offer her agent any incentive to become informed. We formalize this idea in Proposition 7.

PROPOSITION 7. Suppose that $\theta \in (0.5,1)$, $q \in (1 - 1/2\theta,1]$, and $u(\cdot)$ is a Von Neumann-Morgenstern utility function $u(\cdot)$ satisfying $u' > 0, u'' \leq 0$.

(i) Suppose that $u(1) \leq \mu^*(\theta,q)u(R) + (1 - \mu^*(\theta,q))u(0)$. Then there exists a function $g(\cdot)$ such that $g' > 0, g'' \leq 0$, and $g(u(1)) \geq \mu^*(\theta,q)g(u(R)) + (1 - \mu^*(\theta,q))g(u(0))$.

(ii) Suppose that $u(1) \geq \mu^*(\theta,q)u(R) + (1 - \mu^*(\theta,q))u(0)$. Then for any function $g(\cdot)$ such that $g' > 0$, $g'' \leq 0$, $g(u(1)) \geq \mu^*(\theta,q)g(u(R)) + (1 - \mu^*(\theta,q))g(u(0))$.

(iii) In both (i) and (ii), $v(\cdot) = g(u(\cdot))$ is a Von Neumann-Morgenstern utility function that represents preferences that are globally more risk-averse than those represented by $u(\cdot)$.

Suppose that Marta, whose preferences are represented by $u(\cdot)$, wishes to give her agent the incentive to become informed about x. The proposition establishes that there exist preferences that are globally more risk-averse than Marta's under which a principal would prefer not to give her agent the incentive to become informed. Furthermore, if Marta does not wish to give her agent an incentive to become informed about the state x, then any principal who is globally more risk-averse than Marta would also choose not to offer incentives to an identically biased agent facing the same investment decision.

The preceding analysis reflects an assumption that a biased agent who feels he is fully informed will invest all of the principal's wealth in a single risky asset. The agent will pursue such a strategy if, for example, the principal offers the agent a fixed share of the principal's gross investment return. Our analysis assumes, of course, that the principal cannot directly contract on decisions,

only on returns. But it also implicitly assumes that the principal cannot punish the agent for having too high an expected return. If she could, then she might wish for the agent to gather some information—and then provide incentives such that the (overconfident) agent will be afraid of making too much money for the principal.[27] There might be a variety of reasons, of course, why such contracts are infeasible or undesirable. If, for example, the principal is uncertain about either the true value of R or the extent of the agent's confirmatory bias q, she may not have enough information to propose a contract that always leads the agent to invest optimally.

Nevertheless, even in the presence of uncertainty the principal would generally be better off if she could restrain the agent's ability to take an extreme action. If feasible to restrain the agent, the principal could propose a contract that stipulates that the agent cannot invest more than a fraction of her wealth in any single asset. Even more simply, the principal could simply give the agent only a portion of her wealth to invest. All of these strategies serve the same purpose, namely preventing an overconfident agent from investing too much of the principal's money in a single asset while still taking advantage of the information that the agent actually does possess.

VI. Discussion and Conclusion

We believe that confirmatory bias is important in many social and economic situations, and that variants of the formulation developed in this paper can be usefully applied in formal economic models. For instance, confirmatory bias is likely to matter when a

27. From the principal's point of view, the optimal proportional allocation to a risky investment, a^*, maximizes the objective function $V(a) \equiv \mu^*(\theta,q)\, u(1 + (R - 1)a) + (1 - \mu^*(\theta,q)) \cdot (1 - a)$. The optimal allocation a^* then satisfies the following necessary and sufficient condition:

$$\begin{aligned} &\geq 0, && a^* = 1 \\ \mu^*(\theta,q)u'(1 + (R - 1)a^*)(R - 1) - (1 - \mu^*(\theta,q))u'(1 - a^*) = 0, &&\quad a^* \in [0,1] \\ &\leq 0, && a^* = 0. \end{aligned}$$

The principal would like to propose a contract specifying that the agent receives an arbitrarily small reward when the principal's gross return is $1 + (R - 1)a^*$, no payoff when the principal's gross return is $1 - a^*$, and a large penalty for any other gross return. Such a contract would punish the agent if he chooses an allocation that is more extreme than the principal desires.

decision-maker must aggregate information from many sources. In a setting where several individuals (nonstrategically) transmit their beliefs to a principal, how should she combine these reports to form her own beliefs? If the principal thought that the agents were Bayesians, then she would be very sensitive to the strength of the agents' beliefs. Suppose, for instance, that the principal knows that all agents receive signals of strength $\theta = .6$. Then if two agents report believing Hypothesis A with probability .6 and one agent reports believing Hypothesis B with probability .77 (meaning he has gotten three more b signals than a signals), the principal should believe in Hypothesis B with probability .6.

What if the principal were aware that agents were subject to confirmatory bias? If confirmatory bias is so severe that only an agent's first signal is very informative, then the principal may wish to discount the *strength* of agents' beliefs and basically aggregate according to a "majority rules" criterion. In the example above, for instance, the principal should perhaps think Hypothesis A is more likely, because two of three agents believe in it. We think this intuition has merit, but it is complicated by the fact that agents who believe relatively weakly in a hypothesis may be more likely to be wrong than right. So, if the principal thought confirmatory bias were severe *and* were very sure that all agents had received lots of information, then in our example she should believe that *all three* agents have provided evidence in favor of Hypothesis B. Hence, she should believe *more* in Hypothesis B than she would if the agents were Bayesian.

We suspect nonetheless that the "majority-rules intuition" is more valid, especially when considering realistic uncertainty by the principal about how many signals each agent has received. If she were highly uncertain about how much information each agent received, she would assume weak beliefs merely reflected that an agent got few signals. Similarly, if the principal thinks susceptibility to confirmatory bias is heterogeneous, she might infer that an agent's weak beliefs indicate merely that he is not susceptible to overconfidence, and count weak beliefs as much as strong beliefs. Indeed, she may then count them *more* heavily, since confirmation-free agents are not only less likely to be overconfident, they are also less likely to be wrong.

This intuition that, when aggregating information from a group, it may be wise to count the number of people with given

beliefs rather than the strength of their convictions suggests a related prescription for organizational design: relative to what she would do with Bayesian agents, a principal may prefer to hire more agents to collect a given amount of information. That is, while the lower value of information processing by confirmatory agents may mean that either more or fewer should be hired than if they were Bayesian, fixing the total amount of information processing a principal wants done, with confirmatory agents she should prefer more people thinking than if they were fully rational agents.

Imagine, for instance, that a principal allocated 1000 "signals" among different agents, whose reports she would aggregate to form her own beliefs. There are various costs that might influence how many agents to have, or (equivalently) how many signals to allocate per agent, e.g., the fixed cost of hiring new agents and decreasing returns from each individual due to fatigue or the increasing opportunity cost of time. But the optimal number of confirmatory agents is likely to be greater than the optimal number of Bayesian agents. Intuitively, the value of allocating a signal to a confirmatory agent is less than the value of allocating it to a Bayesian agent, *unless* the confirmatory agent is unbiased by previous signals. For example, if the principal hires 1000 confirmatory agents, each to report his observation of a single signal, then she receives all of the information contained in the signals. If the principal instead hires one confirmatory agent to report his beliefs after interpreting 1000 signals, she may get far less information. Both signal allocations would yield the same amount of information if the agents were Bayesian.

We suspect that a similar issue plays out less abstractly in different aspects of the legal system. While other explanations are probably more important, confirmatory bias may help to explain some features of the American jury system, such as the bias toward more rather than fewer jurors and the use of a majority-rules criterion with no mechanism (other than jury deliberations) to extract the strength of all participants' convictions. Confirmatory bias may also help to justify the use of multiple judges to reach a decision when using a single judge seems to be more cost-effective. Appeals, for example, are usually heard by a panel of judges that does not include the trial judge, and some legal scholars (e.g., Resnik [1982]) argue that the judge who adjudicates at trial should not also supervise settlement bargaining and

pretrial discovery, which is the process by which litigants request information from each other. These observers fear that the trial judge might learn things during pretrial activities that would "bias" her during the trial. The notion that the quality of the judge's decisions during the trial suffers if she has more information relevant to the case is somewhat puzzling; worries that she can be "biased" by more information certainly flies in the face of the Bayesian model. While there are various types of bias that one could imagine (e.g., that the judge will use her rulings during the trial to punish perceived misbehavior during the discovery process), the evidence on confirmatory bias raises the possibility that the judge will form preconceptions during the discovery phase of litigation that will cause her to misread additional evidence presented at trial.

Finally, the discovery process itself nicely illustrates how the polarization associated with confirmatory bias may have important implications. Discovery takes place when potential litigants think that a trial is relatively likely, and hence wish to engage in the costly effort of preparing for that trial. Nonetheless, litigants often settle their case out of court during or after the discovery process. Discovery encourages this settlement by promoting the exchange of information between the litigants and, hence, helping to align their perceptions of the likely outcome at trial. But while the evidence garnered during the discovery process sometimes does lead to settlement before trial, confirmatory bias suggests that the discovery process may be less efficient at achieving such settlement than would be hoped: if a piece of evidence is ambiguous, it may move the parties' beliefs farther apart. Each litigant will interpret the evidence through the prism of his or her own beliefs, and each may conclude that the evidence supports his or her case. More generally, efforts to reduce disagreements by providing evidence to the parties involved in a conflict may not be as easy to achieve as one would hope.

Much of our discussion above implicitly makes an assumption about judgment whose psychological validity has not (to our knowledge) been determined by research: that somebody designing an institution is aware of the bias of others. We suspect that usefully incorporating confirmatory bias into economic analysis will depend upon the extent to which people believe that others suffer from confirmatory bias. It could be that people are well aware of biases in others' judgment, or that people are unaware of

the general tendency toward confirmatory bias.[28] Investors who hire a money manager might or might not believe that the money manager suffers from a confirmatory bias (and is therefore prone toward overconfidence). A principal hiring an employee to make decisions might or might not know that the employee will be prone to making such errors. By the logic of economic models that involve multiple agents, these distinctions are likely to matter: Just as assuming that rationality is common knowledge is often very different than merely assuming that people are rational, assuming that agents are aware of others' irrationality may be very different than merely assuming that people are irrational.

How might economic implications depend on people's awareness of others' confirmatory bias? One possibility is that people might exploit the bias of others. A principal may, for instance, design an incentive contract for an agent that yields the agent lower wages on average than the agent anticipates, because the agent will be overconfident about her judgments in ways that may lead her to exaggerate her yield from a contract. Conversely, others may wish to mitigate bias rather than exploit it. A principal may be more concerned with overcoming costly bias of an agent than with exploiting it, and design contracts that avoid errors.

Appendix 1: Differential-Strength Signals and Underconfidence

If the agent receives signals of different strengths in different periods, it is possible that the agent will be *underconfident* in his belief about which of the two states is most likely. Suppose, for example, that the agent receives three signals $s_t \in \{a,b\}$, $t \in \{1,2,3\}$. Suppose that the first two signals are distributed according to prob $(s_t = a|A) =$ prob $(s_t = b|B) = \theta > 0.5$, $t \in \{1,2\}$, but that the agent's third signal is distributed according to prob $(s_3 = a|A) =$ prob $(s_3 = b|B) = \theta^3/[\theta^3 + (1 - \theta)^3]$. That is, the agent's third signal is three times as strong as first- or second-period signals. As before, with probability $q > 0$ the agent misreads signals that conflict with his belief about which state is more likely. (This

28. Unfortunately, while this issue may turn out to be central to economic applications of confirmatory bias (and to applications of other psychological biases), we have not found psychological research that convincingly resolves this issue. There is a small literature in "construal" that concerns third-party awareness of biases. See, e.g., Ross [1987], and tangentially Paese and Kinnaly [1993]. We have not found investigation of this issue in the context of confirmatory bias or overconfidence.

means that the probability of misreading is independent of the strength of the signal.)

Suppose that the agent perceives that his first two signals support Hypothesis B, while his third signal supports Hypothesis A. Formally, the agent perceives $(\sigma_1 = \beta, \sigma_2 = \beta, \sigma_3 = \alpha)$. Given these perceived signals, the agent's posterior likelihood ratio is $\Lambda(s_1 = b, s_2 = b, s_3 = a) = \theta/(1 - \theta) > 1$. Now, suppose that a Bayesian observer knows both that the agent's posterior likelihood ratio is $\Lambda = \theta/(1 - \theta)$ and that the agent suffers from confirmatory bias. Given the distributions of the signals, the observer is able to infer that the agent has perceived $(\sigma_1 = \beta, \sigma_2 = \beta, \sigma_3 = \alpha)$. Then, the observer's belief regarding the relative likelihood that the state is $x = A$ versus $x = B$ is given by

$$\Lambda^*(b,\beta,\alpha) = \frac{(1 - \theta)(1 - \theta + q\theta)(1 - q)\theta^3}{\theta(\theta + q(1 - \theta))(1 - q)(1 - \theta)^3}$$

$$= \frac{(1 - \theta + q\theta)\theta^2}{(\theta + q(1 - \theta))(1 - \theta)^2} > \frac{\theta}{1 - \theta}, \qquad \forall q \in (0,1].$$

Therefore, given what she infers about the agent's sequence of perceived signals, a Bayesian observer believes that the biased agent is *underconfident* in his belief that the true state is A.

This underconfidence result arises here because the observer infers the exact sequence of the agent's perceived signals from his likelihood ratio. In this light, the results here are the same as the path-dependent underconfidence example in the text—if the agent is known to have only recently come to believe in a hypothesis, then he will be underconfident. In our main model, in which the agent receives signals of equal strength, an observer who knows the agent's beliefs cannot infer the exact sequence of the agent's perceived signals.

While there may be some domains in which this differential-signal model is applicable, constructing examples of underconfidence seem to require clever contrivance. It is first of all clear that the "overconfidence" result will be stronger than the underconfidence result in one sense: in the model of this paper, the overconfidence result holds for *all* final beliefs by the agent. Any underconfidence example will clearly hold for only *some* final beliefs—because it will always be the case that a confirmatory agent is overconfident when all his perceived signals favor one hypothesis.

We suspect, moreover, that more complicated and weaker versions of Proposition 1 will hold in more general models. The underconfidence result seems to rely on the agent having received a small number of signals, where certain final beliefs can only be generated by a unique path of updating. Consequently, it is very likely that a "limit overconfidence" result would hold—once an agent is likely to have received large numbers of signals of all strengths, we can assure that $\Lambda^* < \Lambda$ when $\Lambda > 1$.

Appendix 2: Proofs

Proof of Proposition 1. We first notice that

$$\text{prob}\,(n_\alpha,n_\beta|A) = \sum_{i=0}^{n_\beta} \text{prob}\,(i,i|A)c(n_\alpha - n_\beta,n_\alpha + n_\beta - 2i) \cdot \theta[\theta + q(1-\theta)]^{n_\alpha-1-i}[(1-q)(1-\theta)]^{n_\beta-i}$$

and

$$\text{prob}\,(n_\alpha,n_\beta|B) = \sum_{i=0}^{n_\beta} \text{prob}\,(i,i|B)c(n_\alpha - n_\beta,n_\alpha + n_\beta - 2i) \cdot (1-\theta)[(1-\theta) + q\theta]^{n_\alpha-1-i}[(1-q)\theta]^{n_\beta-i},$$

where $c(n_\alpha - n_\beta,n_\alpha + n_\beta - 2i)$ is the number of ways to choose $n_\alpha - n_\beta$ more a signals than b signals in $n_\alpha + n_\beta - 2i$ draws without ever having chosen an equal number of a and b signals, and prob $(i,i|x)$ is the probability of observing i perceived a and i perceived b signals in $2i$ draws when the true state is $x \in \{A,B\}$. Given the symmetric distribution of the signals, prob $(i,i|A)$ = prob $(i,i|B)$.[29] Therefore, prob $(n_\alpha,n_\beta|A)$ and prob $(n_\alpha,n_\beta|B)$ differ

29. Formally,

$$\text{prob}\,(i,i|A) = \text{prob}\,(i,i|B) = \sum_{j=0}^{i}\sum_{k=0}^{i-j}\sum_{l=0}^{\max\{i-j-k-1,0\}} \cdot d_{jkl}\theta^j\theta^{**k}((1-q)(1-\theta))^{j+k}(1-\theta)^{i-j-k-l}\theta^{*l}((1-q)\theta)^{i-j-k}.$$

The coefficient d_{jkl} is the number of ways to choose j signals in favor of the correct hypothesis when the agent believes the two hypotheses are equally likely, k biased signals favoring the correct hypothesis, $i-j-k-l$ signals in favor of the incorrect hypothesis when the agent believes the two hypotheses are equally likely, l biased signals in favor of the incorrect hypothesis, $j+k$ unbiased signals opposing a belief in favor of the correct hypothesis, and $i-j-k$ unbiased signals opposing a belief in favor of the incorrect hypothesis.

only by the effect of the signals that the agent perceives after the last time that he believes the two hypotheses are equally likely.

Using Bayes' Rule,

$$\Lambda^*(n_\alpha,n_\beta) = \frac{\text{prob}\,(n_\alpha,n_\beta|A)}{\text{prob}\,(n_\alpha,n_\beta|B)} \tag{1.1}$$

$$= \frac{\begin{array}{r}\Sigma_{i=0}^{n_\beta}\,\text{prob}\,(i,i|A)c(n_\alpha - n_\beta,n_\alpha + n_\beta - 2i)\\ \cdot\,\theta[\theta + q(1-\theta)]^{n_\alpha-1-i}[(1-q)(1-\theta)]^{n_\beta-i}\end{array}}{\begin{array}{r}\Sigma_{i=0}^{n_\beta}\,\text{prob}\,(i,i|B)c(n_\alpha - n_\beta,n_\alpha + n_\beta - 2i)(1-\theta)\\ \cdot\,[(1-\theta) + q\theta]^{n_\alpha-1-i}[(1-q)\theta]^{n_\beta-i}\end{array}}.$$

Because $[\theta + q(1-\theta)]/[(1-\theta) + q\theta] < \theta/(1-\theta)$, $\forall q \in (0,1]$, it follows that

$$[\theta + q(1-\theta)]^{n_\alpha-1-i}(1-\theta)^{n_\alpha-1-i} \le [(1-\theta) + q\theta]^{n_\alpha-1-i}\theta^{n_\alpha-1-i} \tag{1.2}$$

with a strict inequality for $i = 0$ since the hypotheses imply that $n_\alpha \ge 2$. Factoring and multiplying (1.2) by $(1-\theta)(1-q)^{n_\beta-i}$ and rearranging, we have

$$\theta[\theta + q(1-\theta)]^{n_\alpha-1-i}(1-q)^{n_\beta-i}(1-\theta)^{n_\beta-i} \tag{1.3}$$

$$\le (1-\theta)[(1-\theta) + q\theta]^{n_\alpha-1-i}(1-q)^{n_\beta-i}\theta^{n_\beta-i}\left(\frac{\theta}{1-\theta}\right)^{n_\alpha-n_\beta}$$

$\forall i$, with a strict inequality for at least $i = 0$ since $n_\alpha \ge 2$. Using (1.1), (1.3), and prob $(i,i|A)$ = prob $(i,i|B)$,

$$\Lambda^*(n_\alpha,n_\beta) < \frac{\begin{array}{r}\Sigma_{i=0}^{n_\beta}\,\text{prob}\,(i,i|A)c(n_\alpha - n_\beta,n_\alpha + n_\beta - 2i)(1-\theta)\\ \cdot\,[(1-\theta) + q\theta]^{n_\alpha-1-i}(1-q)^{n_\beta-i}\theta^{n_\beta-i}(\theta/(1-\theta))^{n_\alpha-n_\beta}\end{array}}{\begin{array}{r}\Sigma_{i=0}^{n_\beta}\,\text{prob}\,(i,i|B)c(n_\alpha - n_\beta,n_\alpha + n_\beta - 2i)(1-\theta)\\ \cdot\,[(1-\theta) + q\theta]^{n_\alpha-1-i}(1-q)^{n_\beta-i}\theta^{n_\beta-i}\end{array}}$$

$$= \left(\frac{\theta}{1-\theta}\right)^{n_\alpha-n_\beta} = \Lambda(n_\alpha,n_\beta).$$

Proof of Proposition 2. Clearly $\lim_{\epsilon\to 0}\Lambda^*(n_\alpha,0|1-\epsilon,1-\epsilon) = \infty\forall n_\alpha > 0$ and $\lim_{\epsilon\to 0}\Lambda^*(n_\alpha,1|1-\epsilon,1-\epsilon) = (n_\alpha+1)/(n_\alpha-1)\forall \cdot n_\alpha > 1$.

It can be shown that, if the agent's current beliefs are that

A and B are equally likely, and A is true,

then the probability that the next signal is $\alpha \approx 1$ is $\beta = \epsilon$

A and B are equally likely, and B is true,
then the probability that the next signal is $\alpha = \epsilon$ is $\beta \approx 1$
A is probably true, and A is true,
then the probability that the next signal is $\alpha \approx 1$ is $\beta = \epsilon^2$
A is probably true, and B is true,
then the probability that the next signal is $\alpha \approx 1$ is $\beta \approx \epsilon$
B is probably true, and A is true,
then the probability that the next signal is $\alpha \approx \epsilon$ is $\beta \approx 1$
B is probably true, and B is true,
then the probability that the next signal is $\alpha = \epsilon^2$ is $\beta \approx 1$.

From these numbers we can calculate that, if $n_\alpha > n_\beta$,

- Suppose that A is the true state. Consider all paths σ^* such that (1) $\sigma_1^* = \beta$ and (2) there is always a strict majority of β signals until $2n_\beta - 1$ signals, after which all signals are α. Then the probability of any particular path σ^* is about $\epsilon^{n_\beta+1}$. All other paths each occur with probability on the order of $\epsilon^{n_\beta+2}$ or greater when $n_\beta \geq 2$.
- Suppose that B is the true state. Consider all paths σ^{**} such that (1) $\sigma_1^{**} = \alpha$ and (2) there is always a strict majority of α signals. The probability of any particular path σ^{**} is about $\epsilon^{n_\beta+1}$. All other paths each occur with probability on the order of $e^{n_\beta+2}$ or greater when $n_\beta \geq 2$.

To show that $\Lambda^*(n_\alpha,n_\beta|1-\epsilon,1-\epsilon) < 1$ with $n_\beta \geq 2$, therefore, we need only to show that the number of paths of type σ^{**} is strictly greater than the number of paths of type σ^*. This is easy to verify. For every particular path of type σ^*, there exists a path of type σ^{**} that is the mirror image of that path for the first $2n_\beta - 1$ signals (replacing each α with a β and each β with an α), and whose last $n_\alpha - n_\beta + 1$ signals consist of $n_\alpha - n_\beta - 1$ α's followed by 2 β's. In addition, there will exist at least one more path of type σ^{**}; for instance, n_α α's followed by n_β β's.

QED

Proof of Proposition 3. The proof is by induction. Define $\Lambda(n)$ as the agent's relative likelihood ratio after observing n signals. Suppose that $\Lambda(1) > 1$. Then a Bayesian observer infers that the agent observed a single true "a" signal, and $\Lambda^*(1) = \theta/(1-\theta) > 1$. Now suppose that $\Lambda(n)$, $\Lambda^*(n)$, and $\Lambda(n+1) > 1$. We must show that $\Lambda^*(n+1) > 1$. First, suppose that n is an even number. Because $\Lambda(n) > 1$, after period n the agent has perceived at least

two more "a" signals than "b" signals. Therefore, knowing only that $\Lambda(n) > 1$, a Bayesian observer's relative likelihood ratio, $\Lambda^*(n) = \text{prob}\,(x = A)/\text{prob}\,(x = B)$, is given by

$$\Lambda^*(n) = \frac{\begin{array}{r}\Sigma_{j=0}^{(n/2)-1}\,\Sigma_{i=0}^{j}\,p(i,i|A)c(n-2j,n-2i)\\ \cdot\,\theta[\theta+q(1-\theta)]^{n-1-j-i}(1-q)^{j-i}(1-\theta)^{j-i}\end{array}}{\begin{array}{r}\Sigma_{j=0}^{(n/2)-1}\,\Sigma_{i=0}^{j}\,p(i,i|B)c(n-2j,n-2i)(1-\theta)\\ \cdot\,[(1-\theta)+q\theta]^{n-1-j-i}(1-q)^{j-i}\theta^{j-i}\end{array}} = \frac{p^A(n)}{p^B(n)}, \tag{3.1}$$

where $p(i,i|x)$ and $c(\cdot,\cdot)$ are defined as in the proof of Proposition 1. Define $p^x(n)$ as the probability of perceiving a (strict) majority of "α" signals in n draws given the state $x \in \{A,B\}$. Then $\Lambda^*(n+1)$ is given by

$$\Lambda^*(n+1) = \frac{p^A(n) + p(n/2,n/2|A)\theta}{p^B(n) + p(n/2,n/2|B)(1-\theta)}. \tag{3.2}$$

Because $p(n/2,n/2|A) = p(n/2,n/2|B)$ and $\theta > 0.5$, $\Lambda^*(n+1) > 1$ follows immediately from the hypothesis that $\Lambda^*(n) > 1$, which implies that $p^A(n) > p^B(n)$.

Now suppose that n is an odd number. Because by hypothesis $\Lambda(n) > 1$, after period n the agent has perceived more "a" than "b" signals. Therefore, knowing only that $\Lambda(n) > 1$, a Bayesian observer's relative likelihood ratio, $\Lambda^*(n) = \text{prob}\,(x = A)/\text{prob}\,\cdot\,(x = B)$, is given by

$$\Lambda^*(n) = \frac{\begin{array}{r}\Sigma_{j=0}^{(n-1)/2}\,\Sigma_{i=0}^{j}\,p(i,i|A)c(n-2j,n-2i)\\ \cdot\,\theta[\theta+q(1-\theta)]^{n-1-j-i}(1-q)^{j-i}(1-\theta)^{j-i}\end{array}}{\begin{array}{r}\Sigma_{j=0}^{(n-1)/2}\,\Sigma_{i=0}^{j}\,p(i,i|B)c(n-2j,n-2i)(1-\theta)\\ \cdot\,[(1-\theta)+q\theta]^{n-1-j-i}(1-q)^{j-i}\theta^{j-i}\end{array}} = \frac{p^A(n)}{p^B(n)}. \tag{3.3}$$

Meanwhile, $\Lambda^*(n+1)$ is given by

$$\Lambda^*(n+1) = \frac{\begin{array}{r}p^A(n) - \Sigma_{i=0}^{(n-1)/2}\,p(i,i|A)c(1,n-2i)\\ \cdot\,\theta[\theta+q(1-\theta)]^{((n-1)/2)-i}[(1-q)\\ \cdot\,(1-\theta)]^{((n-1)/2)-i}(1-q)(1-\theta)\end{array}}{\begin{array}{r}p^B(n) - \Sigma_{i=0}^{(n-1)/2}p(i,i|B)c(1,n-2i)\\ \cdot\,(1-\theta)[(1-\theta)+q\theta]^{((n-1)/2)-i}\\ \cdot\,[(1-q)\theta]^{((n-1)/2)-i}(1-q)\theta\end{array}}. \tag{3.4}$$

Because $\Lambda^*(n) > 1$ implies that $p^A(n) > p^B(n)$, in order to establish $\Lambda^*(n+1) > 1$ it is sufficient to show that

$$\sum_{i=0}^{(n-1)/2} p(i,i|A)c(1,n-2i)\theta[\theta + q(1-\theta)]^{((n-1)/2)-i}[(1-q) \cdot (1-\theta)]^{((n-1)/2)-i}(1-q)(1-\theta) \leq \sum_{i=0}^{(n-1)/2} p(i,i|B)c(1,n-2i)(1-\theta)[(1-\theta) + q\theta]^{((n-1)/2)-i}[(1-q)\theta]^{((n-1)/2)-i}(1-q)\theta. \tag{3.5}$$

Using the fact that $p(i,i,|A) = p(i,i|B)$ and canceling like terms, the inequality in (3.5) is satisfied if

$$[\theta + q(1-\theta)]^{((n-1)/2)-i}(1-\theta)^{((n-1)/2)-i} \leq [1-\theta+q\theta]^{((n-i)/2)-1}\theta^{((n-1)/2)-i} \qquad \forall i \in \{0, \ldots, (n-1)/2\}.$$

But this inequality is always satisfied because $(\theta + q(1-\theta))/((1-\theta)+q\theta) < \theta/(1-\theta) \forall q \in (0,1]$. Therefore, $\Lambda^*(n+1) > 1$.

QED

Proof of Proposition 4. The first hypothesis implies that $\theta^* > 0.5$, and therefore, using Lemma 1, P_W satisfies

$$P_W = (1-\theta) \cdot [(1 - p(1,\theta^*)) + p(1,\theta^*) \cdot P_W] + \theta \cdot [p(1,\theta^{**}) \cdot P_W],$$

or

$$P_W = \frac{(1-\theta) \cdot (1 - ((1-\theta^*)/\theta^*))}{(1-(1-\theta)\cdot((1-\theta^*)/\theta^*) - \theta((1-\theta^{**})/\theta^{**}))}.$$

$P_W > 0$ because $\theta^{**} > 0.5$ for all $q \geq 0$.

The second hypothesis implies that $\theta^* \leq 0.5$, and therefore, using Lemma 1, P_W satisfies

$$P_W = (1-\theta) \cdot P_W + \theta \cdot [p(1,\theta^{**}) \cdot P_W].$$

$P_W = 0$ because $p(1,\theta^{**}) < 1$.

QED

Proof of Proposition 5. Ignoring integer problems, and since $q > 1 - 1/(2\theta)$ for the cases we consider below, the definition of

$D(\mu)$ and Lemma 1 imply that

$$P_W(\mu) = \left(1 - \left[\frac{1-\theta^*}{\theta^*}\right]^{D(\mu)}\right) + \left[\frac{1-\theta^*}{\theta^*}\right]^{D(\mu)} P_W(0.5) > 0.$$

(i) Note that $\lim_{q\to 1} \theta^* = 1$ for all (μ,θ). Therefore, for q sufficiently close to 1, $P_W(\mu)$ can be made arbitrarily close to 1.

(ii) Note that $\lim_{\theta\to 0.5} \theta^* > 0.5$ and $\lim_{\theta\to 0.5} D(\mu) = \infty$ for all (μ,q). Therefore, for θ sufficiently close to 0.5, $P_W(\mu)$ can be made arbitrarily close to 1.

QED

Proof of Proposition 6. (i) Fix $\theta \in (0.5,\underline{\mu}\,]$. The result follows directly from the fact that the principal's payoff from having the agent observe signals, $\Pi(\theta,q) = \mu^*(\theta,q)u(WR) + (1 - \mu^*(\theta,q))u(0)$, is continuously monotone decreasing in q, with $\Pi(\theta,1 - 1/2\theta) > u(W)$ and $\Pi(\theta,1) \leq u(W)$. (ii) Fix $q \in (0,1]$. The result follows directly from the fact that $\Pi(\theta,q)$ is continuously monotone increasing in θ, with $\Pi(0.5,q) \leq u(W)$ and $\Pi(u,q) > u(W)$.

QED

Proof of Proposition 7. The proof is by construction. In order to establish the result, it is sufficient to show that there exists a function $g(\cdot)$ such that

$$\frac{g(u(WR) - g(u(W))}{g(u(W)) - g(u(0))} \leq \frac{1 - \mu^*(\theta,q)}{\mu^*(\theta,q)}.$$

Define the function $g(\cdot)$ as

$$g(x) = \begin{cases} x, x \leq u(W) \\ u(W) + \epsilon(x - u(W)), x > u(W). \end{cases}$$

Clearly, $g' > 0, g'' \leq 0$, and

$$\frac{g(u(WR) - g(u(W))}{g(u(W)) - g(u(0))} = \frac{\epsilon(u(WR) - u(W))}{u(W) - u(0)} \leq \frac{1 - \mu^*(\theta, q)}{\mu^*(\theta, q)}$$

for ϵ sufficiently small. Parts (ii) and (iii) follow directly from concavity of $g(\cdot)$ and Theorem 1 in Pratt [1964].

QED

UNIVERSITY OF CALIFORNIA, BERKELEY
EMORY UNIVERSITY

REFERENCES

Arkes, H. R., "Principles in Judgment/Decision Making Research Pertinent to Legal Proceedings," *Behavioral Sciences and the Law,* VII (1989), 429–456.
Bacon, F. (1620), *The New Organon and Related Writings* (New York, NY: Liberal Arts Press, 1960).
Baumann, A. O., R. B. Deber, and G. G. Thompson, "Overconfidence among Physicians and Nurses: The 'Micro-Certainty, Macro-Uncertainty,' Phenomenon," *Social Science and Medicine,* XXXII (1991), 167–174.
Beattie, J., and J. Baron, "Confirmation and Matching Biases in Hypothesis Testing," *Quarterly Journal of Experimental Psychology: Human Experimental Psychology,* XL (1988), 269–297.
Bjorkman, M., "Internal Cue Theory: Calibration and Resolution of Confidence in General Knowledge," *Organizational Behavior and Human Decision Processes,* LVIII (1994), 386–405.
Bodenhausen, G. V., and M. Lichtenstein, "Social Stereotypes and Information-Processing Strategies: The Impact of Task Complexity," *Journal of Personality and Social Psychology,* LII (1987), 871–880.
Bodenhausen, G. V., and R. S. Wyer, "Effects of Stereotypes in Decision Making and Information-Processing Strategies," *Journal of Personality and Social Psychology,* XLII (1985), 267–282.
Borum, R., R. Otto, and S. Golding, "Improving Clinical Judgment and Decision Making in Forensic Evaluation," *Journal of Psychiatry and Law,* XXI (1993), 35–76.
Bruner, J., and M. Potter, "Inference in Visual Recognition," *Science,* CXLIV (1964), 424–425.
Camerer, C., "Individual Decision Making," in *Handbook of Experimental Economics,* J. Kagel and A. E. Roth, eds. (Princeton, NJ: Princeton University Press, 1995), pp. 587–703.
Chapman, L. J., and J. P. Chapman, "Genesis of Popular but Erroneous Psychodiagnostic Observations," *Journal of Abnormal Psychology,* LXXII (1967), 193–204.
Chapman, L. J., and J. P. Chapman, "Illusory Correlation as an Obstacle to the Use of Valid Psychodiagnostic Signs," *Journal of Abnormal Psychology,* LXXIV (1969), 271–280.
Chapman, L. J., and J. P. Chapman, "Test Results Are What You Think They Are," *Psychology Today,* V (1971), 106.
Darley, J., and P. Gross, "A Hypothesis-Confirming Bias in Labeling Effects," *Journal of Personality and Social Psychology,* XLIV (1983), 20–33.
Devine, P. G., E. R. Hirt, and E. M. Gehrke, "Diagnostic and Confirmation Strategies in Trait Hypothesis Testing," *Journal of Personality and Social Psychology,* LVIII (1990), 952–963.
Dougherty, T. W., D. B. Turban, and J. C. Callender, "Confirming First Impressions in the Employment Interview: A Field Study of Interviewer Behavior," *Journal of Applied Psychology,* LXXIX (1994), 659–665.
Einhorn, H., and R. Hogarth, "Confidence in Judgment: Persistence of the Illusion of Validity," *Psychological Review,* LXXXV (1978), 395–416.
Feller, W., *An Introduction to Probability Theory and Its Applications* (New York, NY: John Wiley and Sons, Inc., 1968).
Fischoff, B., and R. Beyth-Marom, "Hypothesis Evaluation from a Bayesian Perspective," *Psychological Review,* XC (1983), 239–260.
Fischhoff, B., P. Slovic, and S. Lichtenstein, "Knowing with Certainty: The Appropriateness of Extreme Confidence," *Journal of Experimental Psychology: Human Perception and Performance,* III (1977), 552–564.
Fleming, J., and A. J. Arrowood, "Information Processing and the Perseverance of Discredited Self-Perceptions," *Personality and Social Psychology Bulletin,* V (1979), 201–205.
Friedrich, J., "Primary Error Detection and Minimization (PEDMIN) Strategies in Social Cognition: A Reinterpretation of Confirmation Bias Phenomena," *Psychological Review,* C (1993), 298–319.
Griffin, D., and A. Tversky, "The Weighing of Evidence and the Determinants of Confidence," *Cognitive Psychology,* XXIV (1992), 411–435.

Hamilton, D. L., and T. L. Rose, "Illusory Correlation and the Maintenance of Stereotypic Beliefs," *Journal of Personality and Social Psychology,* XXXIX (1980), 832–845.

Hamilton, D. L., S. J. Sherman, and C. M. Ruvolo, "Stereotype-Based Expectancies: Effects on Information Processing and Social Behavior," *Journal of Social Issues,* XLVI (1990), 35–60.

Haverkamp, B. E., "Confirmatory Bias in Hypothesis Testing for Client-Identified and Counselor Self-Generated Hypotheses," *Journal of Counseling Psychology,* XL (1993), 303–315.

Hodgins, H. S., and M. Zuckerman, "Beyond Selecting Information: Biases in Spontaneous Questions and Resultant Conclusions," *Journal of Experimental Social Psychology,* XXIX (1993), 387–407.

Hubbard, M., "Impression Perseverance: Support for a Cognitive Explanation," *Representative Research in Social Psychology,* XIV (1984), 48–55.

Jennings, D. L., M. R. Lepper, and L. Ross, "Persistence of Impressions of Personal Persuasiveness: Perseverance of Erroneous Self-Assessments outside the Debriefing Paradigm," *Personality and Social Psychology Bulletin,* VII (1981), 257–263.

Jennings, D. L., T. M. Amabile, and L. Ross, "Informal Covariation Assessment: Data-Based versus Theory-Based Judgments," in *Judgment under Uncertainty: Heuristics and Biases,* D. Kahneman, P. Slovic, and A. Tversky, eds. (Cambridge: Cambridge University Press, 1982), pp. 211–230.

Keren, G., "Facing Uncertainty in the Game of Bridge: A Calibration Study," *Organizational Behavior and Human Decision Processes,* XXXIX (1987), 98–114.

——, "On the Ability of Monitoring Non-Veridical Perceptions and Uncertain Knowledge: Some Calibration Studies," *Acta Psychologica,* LXVII (1988), 95–119.

Klayman, J., and Y.-w. Ha, "Confirmation, Disconfirmation, and Information in Hypothesis Testing," *Psychological Review,* XCIV (1987), 211–228.

Lepper, M. R., L. Ross, and R. R. Lau, "Persistence of Inaccurate Beliefs about the Self: Perseverance Effects in the Classroom," *Journal of Personality and Social Psychology,* L (1986), 482–491.

Lord, C. G., L. Ross, and M. R. Lepper, "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence," *Journal of Personality and Social Psychology,* XXXVII (1979), 2098–2109.

Macan, T. H., and R. L. Dipboye, "The Effects of the Application on Processing of Information from the Employment Interview," *Journal of Applied Social Psychology,* XXIV (1994), 1291–1314.

Mahajan, J., "The Overconfidence Effect in Marketing Management Predictions," *Journal of Marketing Research,* XXIX (1992), 329–342.

Mahoney, M., "Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System," *Cognitive Therapy and Research,* I (1977), 161–175.

Mehle, T., C. F. Gettys, C. Manning, S. Baca, and S. Fisher, "The Availability Explanation of Excessive Plausibility Assessments," *Acta Psychologica,* XLIX (1981), 127–140.

Miller, A. G., J. W. McHoskey, C. M. Bane, and T. G. Dowd, "The Attitude Polarization Phenomenon: Role of Response Measure, Attitude Extremity, and Behavioral Consequences of Reported Attitude Change," *Journal of Personality and Social Psychology,* LXIV (1993), 561–574.

Nisbett, R., and L. Ross, *Human Inference: Strategies and Shortcomings of Social Judgment* (Englewood Cliffs, NJ: Prentice-Hall, 1980).

O'Donoghue, E., and M. Rabin, "Incentives for Procrastinators," Northwestern University, CMSEMS Discussion Paper No. 1181, February 25, 1997.

Oskamp, S., "Overconfidence in Case-Study Judgments," in *Judgment under Uncertainty: Heuristics and Biases,* D. Kahneman, P. Slovic, and A. Tversky, eds. (Cambridge: Cambridge University Press, 1982), pp. 287–293.

Paese, P. W., and M. Kinnaly, "Peer Input and Revised Judgment: Exploring the Effects of (Un)Biased Confidence," *Journal of Applied Social Psychology,* XXIII (1993), 1989–2011.

Perkins, D. N., *The Mind's Best Work* (Cambridge, MA: Harvard University Press, 1981).
Pfeifer, P. E., "Are We Overconfident in the Belief That Probability Forecasters Are Overconfident?" *Organizational Behavior and Human Decision Processes,* LVIII (1994), 203–213.
Plous, S., "Biases in the Assimilation of Technological Breakdowns: Do Accidents Make Us Safer?" *Journal of Applied Social Psychology,* XXI (1991), 1058–1082.
Popper, Karl, *Conjectures and Refutations* (London: Routledge and Kegan Paul, Ltd., 1963).
Pratt, J., "Risk Aversion in the Small and in the Large," *Econometrica,* XXXII (1964), 122–136.
Redelmeier, D., and A. Tversky, "On the Belief That Arthritis Pain is Related to the Weather," *Proceedings of the National Academy of Sciences USA,* XCIII (1996), 2895–2896.
Resnik, J., "Managerial Judges," *Harvard Law Review,* XCVI (1982), 374–445.
Ross, L., "The Problem of Construal in Social Inference and Social Psychology," in *A Distinctive Approach to Psychological Research: The Influence of Stanley Schachter,* R. E. N. Neil, E. Grunberg, Judith Rodin, and Jerome E. Singer, eds. (Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1987), pp. 118–150.
Soll, J., "Determinants of Overconfidence and Miscalibration: The Roles of Random Error and Ecological Structure," *Organizational Behavior and Human Decision Processes,* LXV (1996), 117–137.
Souter, G., "Overconfidence Causes Reinsurers to Underprice Coverages: Consultant," *Business Insurance,* XXVII (1993), 47.
Stangor, C., "Stereotype Accessibility and Information Processing," *Personality and Social Psychology Bulletin,* XIV (1988), 694–708.
Stangor, C., and D. N. Ruble, "Strength of Expectancies and Memory for Social Information: What We Remember Depends on How Much We Know," *Journal of Experimental Social Psychology,* XXV (1989), 18–35.
Tomassini, L. A., I. Solomon, M. B. Romney, and J. L. Krogstad, "Calibration of Auditors' Probabilistic Judgments: Some Empirical Evidence," *Organizational Behavior and Human Performance,* XXX (1982), 391–406.
Van Lenthe, J., "ELI: An Interactive Elicitation Technique for Subjective Probability Distributions," *Organizational Behavior and Human Decision Processes,* LV (1993), 379–413.
Winman, A., and P. Juslin, "Calibration of Sensory and Cognitive Judgments: Two Different Accounts," *Scandinavian Journal of Psychology,* XXXIV (1993), 135–148.
Wood, A. S., "Fatal Attractions for Money Managers," *Financial Analysts Journal,* XLV (1989), 3–5.
Wyatt, D., and D. Campbell, "On the Liability or Stereotype of Hypothesis," *Journal of Abnormal and Social Psychology,* XLVI (1951), 496–500.
Zuckerman, M., C. R. Knee, H. S. Hodgins, and K. Miyake, "Hypothesis Confirmation: The Joint Effect of Positive Test Strategy and Acquiescence Response Set," *Journal of Personality and Social Psychology,* LXVIII (1995), 52–60.