Banker Compensation and Confirmation Bias[1]

Hamid Sabourian, University of Cambridge[2]

Anne C. Sibert, Birkbeck, University of London and CEPR[3]

24 Mar 2009

[1]We are grateful to Willem Buiter and Mike Oaksford for helpful comments. Financial support from the ESRC (grant number RES-156-25-0023) is gratefully acknowledged.

[2]Address: Faculty of Economics, Sidgwick Avenue, Cambridge CB3 9DD, UK; Email: hamid.sabourian@econ.cam.ac.uk

[3]Address: School of Economics, Mathematics and Statistics, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK; Email: asibert@econ.bbk.ac.uk.

Abstract

Confirmation bias refers to cognitive errors that bias one towards one's own prior beliefs. A vast empirical literature documents its existence and psychologists identify it as one of the most problematic aspects of human reasoning. In this paper, we present three related scenarios where rational behaviour leads to outcomes that are observationally equivalent to different types of conformation bias. As an application, the model provides an explanation for how the reward structure in the financial services industry led to the seemingly irrational behaviour of bankers and other employees of financial institutions prior to the credit crisis of that erupted in the summer of 2007.
JEL classification: D81, D82, D83, G21
Key words: confirmation bias, belief persistence, overconfidence, signalling, credit crisis.

1 Introduction

Confirmation bias is a term used by psychologists to refer to a number of apparent cognitive errors involving biases towards one's prior beliefs. Examples include a failure to put sufficient weight on evidence that contradicts one's initial hypothesis, overconfidence in one's own ideas and a tendency to avoid searching for evidence that would disprove one's own theories. Psychologists attribute confirmation bias to several factors. These include emotional reasons, such as embarrassment, stubbornness and hope, and cultural reasons, such as superstition and tradition. They also suggest physiological explanations: it is argued that the evolutionary development of the human brain has facilitated the ability to use heuristics which provide good judgements quickly, but which can also lead to systematic biases. In addition, recent research supports the theory that the human brain arrives at outcomes that promote positive and minimise negative emotional responses.[1]

The purpose of this paper is to demonstrate that in many instances what seems to be confirmation bias may instead be rational behaviour. We consider a scenario where an individual takes an action or makes a decision and the consequences of that action or decision are not known until some time in the future. The individual cares, not just about making the best choice, but also about how competent he is perceived to be in the period between when he acts and when the consequences of his action are revealed. It is demonstrated that the individual's incentive to manipulate beliefs about his ability leads him to distort his actions in a way that is observationally equivalent to confirmation bias.

The above scenario covers a wide range of situations in which confirmation bias appears to occur. For concreteness, we consider a specific example: the behaviour of employees of financial institutions which contributed to the credit crisis that erupted in the summer of 2007. Bankers gambled on a continuation of the US housing boom long after most economists predicted its demise; they were overly optimistic about sustainable leverage ratios; managers of insurance companies and pension funds purchased collateralised

[1] See Westen (2006) et al.

debt obligations and asset-backed securities and appear to have deliberately avoided investigating the details of the underlying assets.[2]

Analyses of the roots of the financial crisis frequently place much of the blame for bankers' behaviour on the rewards system in the financial services industry. Pay is based on a bonus system that depends on perceived talents, rather than on long-term results. Bankers who are viewed as exceptionally talented receive vast rewards, lest they be snatched away by competitors; those viewed as less able quickly find themselves unemployed. Apparently, as a consequence of this system, bankers have an incentive to distort their behaviour and to act in a way that - somehow - makes them look competent, even though it leads to bad results in the long run. Discussing bankers' avid participation in the subprime mortgage market, Allan Meltzer remarked, "These are my MBA students, not just mine but MBAs from Harvard, Stanford, Pennsylvania. They were buying and selling this garbage. Are they so stupid? They got compensated for doing it. If they didn't do it they'd lose their jobs."[3]

We present three variants of a simple model where an expert, who may be thought of as a banker, chooses an observable action. Experts differ in their ability to make the correct decision and this ability is their private information. We model this by assuming that prior to making his decision the expert receives a noisy, but informative, signal indicating which action is best. The probability that the signal is correct is viewed as the expert's competency and it and the signal are known only to the expert. In the long run, it is learned whether the action chosen by the expert is the best one or not and at this later time the expert receives a payoff that is higher if he chose correctly than if he did not. In the short run, however, it is not known if the expert made the right choice and his reward depends instead upon how competent he is perceived to be.

In the first variant of the model, the expert chooses an action and is then confronted with publicly observable conflicting information of known quality. He must then choose

[2] See Rajan (2008) for a discussion of this.

[3] Quoted in Samuelson (2008).

whether or not to change his course of action. We show that relatively able experts, those whose signals are of better quality than the public signal, maintain their original choice as do some or all of the experts whose signals are of lower quality than the public signal. The payoff to masquerading as a more competent expert exceeds the benefit of making a choice that is more likely to be correct. Thus, the payoff structure leads to behaviour which looks like the type of confirmation bias that is known as *belief persistence*.

In the second variant of the model, the expert receives his private information and chooses an action. He is then asked how likely he thinks it is that he is correct. In the long run, if his action turns out to be wrong, then he bears a cost that is increasing in the likelihood that he said that he was correct. Even in this set up, where it is potentially costly to the expert to claim to be correct with high probability and where there is no intrinsic benefit to doing so, if the payoff to being viewed as competent is high enough, experts will claim to be more certain than they are. Relatively competent experts will pool, all claiming to be correct with certainty. Less competent experts will separate, but they too all overstate their ability. This closely resembles the type of confirmation bias known as *overconfidence*.

In the third variant of the model, the expert receives his private information and chooses an action. He is then given the opportunity to acquire costly additional information which, if his initial choice is incorrect, might confirm this. The expert can then choose whether or not to pursue his initial action. In this scenario, relatively competent experts value the additional information less than less competent experts because they are unlikely to be wrong and, hence, unlikely to learn anything. If the most competent expert chooses not to acquire the additional information, then, whether or not acquiring the additional information is observable, a range of less competent experts will pool with the most competent experts and also choose not to acquire more information. This is similar to the classical form of confirmation bias: a tendency to fail to search for disconfirming evidence.

Section 2 contains the model of belief persistence. In section 3, we present a model

of overconfidence. In section 4 we model experts who fail to look for information that might disconfirm their hypotheses. Section 5 is the conclusion.

2 Belief Persistence

"And I think 2000 will be a good year as well." Abby Cohen, famously bullish partner at Goldman Sachs, 1999[4]

Psychologist Raymond Nickerson (1998) comments that the empirical evidence supports the view "that once one has taken a position on an issue, one's primary purpose becomes that of defending or justifying that position." An example of this evidence is a study by Anderson et al (1980) who showed that people cling to beliefs, even when the evidence on which their beliefs is based is weak and is entirely discredited. Their subjects, Stanford undergraduates, were each individually told either that weak statistical evidence confirmed that risk aversion and excellence as a firefighter were positively or negatively correlated. They were then asked to provide an explanation of this result. Afterwards, the researchers admitted that the evidence was bogus. The students were asked what they thought the truth really was. It was found that students who had been told the correlation was positive clung to this belief; those who had been told it was negative clung to that belief. Indeed, a number of students on both sides later commented to the researchers that they did not think the experiment would be a success – no one would believe the opposite hypothesis.

In this section, we present a model of optimising experts who cling to their beliefs in the face of contradictory evidence. We suppose that an "expert" makes a forecast and is then presented with conflicting evidence, after which he has the opportunity to continue with his initial forecast or to change it. In particular, we have in mind an expert who is an employee of a bank or other financial institution who needs to make a forecast so that his customers or his employer can make the best investment decision.

[4]Quoted in Gilpin (1999)

Eventually, it will be revealed whether or not the banker is correct. While the banker would prefer to be later proven right than wrong, his bonus in the meantime depends upon how competent he is perceived to be. We demonstrate that even if new evidence makes an expert believe that his initial forecast is likely to be incorrect, the desire to be seen as competent may prevent him from revising his prediction. Thus, although he is entirely rational, he exhibits behaviour that is observationally equivalent to the cognitive error of belief persistence.

Formally, we assume that one of two events will occur and that *ex ante*, each of the events is equally likely. Initially, the expert receives a noisy privately observed signal indicating which event will occur. The probability that this signal is correct, denoted by π, is also the private information of the expert and it is common knowledge that it is drawn from a uniform distribution on $[1/2, 1]$. We refer to this probability as the expert's *competency*. After receiving his signal the expert forecasts one of the two events. Then, a publicly observed noisy signal indicates which of the two events will occur. This public signal is correct with known probability $\pi^p \in [1/2, 1)$. After observing the public signal, the expert makes a second forecast, either persisting with his original forecast or changing it. Some time in the future, the event occurs and is observed.

The expert's (discounted) payoff is $\chi\Pi + P$, where Π is the market's assessment of the expert's competency after he has made his second forecast, but before the event occurs, and P is a variable that equals one if the expert's forecast later turns out to be correct and zero otherwise. We refer to Π as the expert's *reputation*.[5] The strictly positive parameter χ is the weight that the expert puts on his reputation relative to his desire to forecast correctly.

The form of the objective function – in particular, the short-run payoff for perceived competency – is taken as given because it appears to mimic real-world objective functions in the banking industry and in many other jobs as well. It is likely that it is a consequence

[5]There is no term for the expert's reputation after his first forecast because this forecast conveys no information about the expert's competency to the market.

of a worker being unable to commit himself to long-run employment in the firm. It is assumed that the expert's competency matters to the firm in other ways than his ability to forecast. This is because, as will be seen in this section, more competent experts do not necessarily make better forecasts.

The equilibrium concept used throughout the paper is the natural one for signaling games: the perfect Bayesian equilibrium concept. In signaling games, the first player is the sender of a signal. He has private information and chooses an action. Here, player one is the expert who has private information about his competency and he either changes or does not change his forecast. Player two, here the market, receives the signal. Player two has prior beliefs about the sender's signal and these prior beliefs are common knowledge. Player one's strategy is a probability distribution over possible actions. Player two makes a conjecture about how player one's strategy depends upon player one's type. Then, after observing player one's action, player two updates his beliefs using Bayes' rule. It is required that player one chooses the strategy that maximises his welfare, taking into account player two's conjecture and how his action his will affect player two's posterior beliefs. Player two's conjecture about player one's strategy must turn out to be correct.[6]

As he has no other information and as his priors are flat, in period zero the expert initially forecasts the event that his signal favours. If the public signal favours the same event as his own, then he has no reason to change his forecast. We consider the case where the public signal does *not* favour the same event as his own signal did.

Given that the public information favours an event at odds with the expert's original forecast, the market conjectures that the probability that an expert with competency π does *not* change his forecast is $\Psi_c(\pi)$. The market observes the action A of the expert: either he does not change his forecast ($A = N$) or he does change it ($A = C$) and then the market updates its beliefs about the expert's competency. The market's conjectured joint probability density function of the expert's competency and his action $A = N, C$, conditional on the public signal disagreeing with the expert's original forecast, is denoted

[6] See Fudenberg and Tirole (1992).

by $h(\pi, A)$. The marginal density of A is denoted by $h(A)$. Thus, in accordance with Bayes Rule, the conditional probability density function of π given A and is

$$\begin{aligned} h(\pi|A) &= \frac{h(\pi, A)}{h(A)} = \frac{h(\pi, A)}{\int_{1/2}^{1} h(p, A)\, dp} \\ &= \begin{cases} \frac{\Psi_c(\pi) g(\pi, \pi^p)}{\int_{1/2}^{1} \Psi_c(p) g(p, \pi^p) dp} \text{ if } A = N \text{ and } \int_{1/2}^{1} \Psi_c(p)\, g(p, \pi^p)\, dp > 0 \\ \frac{[1-\Psi_c(\pi)] g(\pi, \pi^p)}{\int_{1/2}^{1} [1-\Psi_c(p)] g(p, \pi^p) dp} \text{ if } A = C \text{ and } \int_{1/2}^{1} [1 - \Psi_c(p)]\, g(p, \pi^p)\, dp > 0, \end{cases} \end{aligned} \tag{1}$$

where $g(\pi, \pi^p)$ is the *ex ante* probability that the public signal differs from the expert's signal when the expert has competency π. We have

$$g(\pi, \pi^p) = \pi(1 - \pi^p) + (1 - \pi)\pi^p. \tag{2}$$

Given that the public information differs from the expert's original forecast, if the market observes action A then its expectation of the expert's competency is

$$\Pi^A = \int_{1/2}^{1} p h(p|A)\, dp. \tag{3}$$

Using Bayes Rule, the probability that the expert attaches to the event that his signal favours occurring is

$$\theta(\pi, \pi^p) \equiv \frac{\pi(1 - \pi^p)}{\pi(1 - \pi^p) + (1 - \pi)\pi^p}. \tag{4}$$

Thus, the expected payoff to an expert with competency π of choosing action A is

$$\begin{cases} \chi \Pi^N + \theta(\pi, \pi^p) \text{ if } A = N \\ \chi \Pi^C + 1 - \theta(\pi, \pi^p) \text{ if } A = C. \end{cases} \tag{5}$$

By equation (5), the expert maximises his payoff if and only if

$$\chi\Pi^N + \theta\left(\pi, \pi^p\right) \left\{ \begin{array}{l} > \\ = \\ < \end{array} \right\} \chi\Pi^C + 1 - \theta\left(\pi, \pi^p\right) \text{ and } \Psi\left(\pi\right) \left\{ \begin{array}{l} = 1 \\ \in [0,1] \\ = 0 \end{array} \right\}, \tag{6}$$

where $\Psi\left(\pi\right)$ is the probability that an expert with competency π does not change his forecast. Experts take $\chi\left(\Pi^N - \Pi^C\right) > 0$ as given and $1 - 2\theta\left(\pi, \pi^p\right)$ is strictly decreasing in π on $\left[\frac{1}{2}, 1\right]$, with $1 - 2\theta\left(\pi^p, \pi^p\right) = 0$. Thus, the experts have a threshold solution to their optimisation problem. If $\chi\left(\Pi^N - \Pi^C\right) > 1 - 2\theta\left(\frac{1}{2}, \pi^p\right)$, then all experts chose strategy N. If there exists a π^* such that $\chi\left(\Pi^N - \Pi^C\right) = 1 - 2\theta\left(\pi^*, \pi^p\right)$, then the type-$\pi$ expert chooses N if $\pi > \pi^*$ and C if $\pi < \pi^*$. If $\pi = \pi^*$ then the type-π expert is indifferent between randomisations over N and C.

In equilibrium the market's conjecture must be consistent: $\Psi_c\left(\pi\right) = \Psi\left(\pi\right)$. Thus, the equilibrium is a threshold equilibrium characterised by a π^* such that all experts with competency at least as high as π^* never change their forecast and all experts with competency lower than π^* always change their forecast. That is,

$$\Psi\left(\pi\right) \left\{ \begin{array}{l} = 1 \\ \in [0,1] \\ = 0 \end{array} \right\} \Leftrightarrow \pi \left\{ \begin{array}{l} > \\ = \\ < \end{array} \right\} \pi^*. \tag{7}$$

Substituting equation (7) into equation (1) and the result into equation (3) yields

$$\Pi^A = \Pi^A\left(\pi^*, \pi^p\right) = \left\{ \begin{array}{l} \int_{\pi^*}^1 pg\left(p, \pi^p\right) dp \,/ \int_{\pi^*}^1 g\left(p, \pi^p\right) dp \text{ if } A = N \text{ and } \pi^* < 1 \\ \int_{1/2}^{\pi^*} pg\left(p, \pi^p\right) dp \,/ \int_{1/2}^{\pi^*} g\left(p, \pi^p\right) dp \text{ if } A = C \text{ and } \pi^* > \frac{1}{2}. \end{array} \right. \tag{8}$$

Assumption 1. $\Pi^N\left(1, \pi^p\right) = \lim_{\pi^* \to 1} \Pi^N\left(\pi^*, \pi^p\right) = 1$ *and* $\Pi^C\left(\frac{1}{2}, \pi^p\right) = \lim_{\pi^* \to 1/2} \Pi^C\left(\pi^*, \pi^p\right) = \frac{1}{2}$.

As seen in equation (1), if a particular action is never chosen in equilibrium, then

Bayes' rule cannot be used to form the posterior distribution if such an action were to be observed. Thus, $\Pi^A(\pi^*, \pi^p)$ is not defined in equation (8) if $A = N$ and $\pi^* = 1$ or if $A = C$ and $\pi^* = \frac{1}{2}$. Thus, if no expert ever changes his forecast, then Bayes' rule cannot be used to specify the market's beliefs, were it to observe the out-of-equilibrium or probability zero phenomenon of an expert changing his forecast. As *any* beliefs are admissible, we make the above intuitively appealing Assumption 1. If all experts change their forecast and the market were to observe an expert not change his forecast (a probability zero event), then the market would believe that the expert had a competency of 1. Likewise, if no expert changes his forecast and the market were to observe an expert change his forecast, the market would believe that the expert had a competency of $\frac{1}{2}$. We discuss the implications of this assumption later in this section.

Using equations (7) and (8), we have the following definition

Definition 1. *An **equilibrium** is a* $\pi^* \in \left[\frac{1}{2}, 1\right]$ *such that*

$$\chi\Pi^N(\pi^*, \pi^p) + \theta(\pi, \pi^p) \left\{ \begin{array}{c} > \\ = \\ < \end{array} \right\} \chi\Pi^C(\pi^*, \pi^p) + 1 - \theta(\pi, \pi^p) \text{ and } \pi^* \left\{ \begin{array}{l} = \frac{1}{2} \\ \in \left[\frac{1}{2}, 1\right] \\ = 1 \end{array} \right\}. \quad (9)$$

Let $\tilde{\chi} := 6(2\pi^p - 1)(3 - 2\pi^p)/(5 - 4\pi^p)$. Then we have the following result.

Proposition 1. *If* $\chi < \tilde{\chi}$*, then there exists a unique equilibrium* π^* *and it has the property that* $\pi^* < \pi^p$*. Furthermore, if* $\chi \geq \tilde{\chi}$ *then there is a unique equilibrium where no expert changes his forecast.*

Proof. See the Appendix.

Proposition 1 demonstrates that if experts care enough about their reputation, then there is a pooling equilibrium where no expert ever changes his forecast when faced with conflicting public information.[7] Otherwise, there is an equilibrium where highly compe-

[7]This pooling equilibrium is similar to the one in Cho and Kreps (1987). The senders of the better-quality signals are not able to separate themselves from the senders of the poorer-quality signals because the action space is limited. It differs from the pooling equilbria in Kreps and Wilson (1982) where the senders of the better quality signals not only have limited ability to separate themselves, but – as they

tent experts (those with $\pi \in [\pi^p, 1]$) do not change their mind in the face of conflicting information because their own information is better than the public information. Experts of intermediate competency (those with $\pi \in (\pi^*, \pi^p)$) also do not change their forecast. Their private information is worse than the public information but the greater-than-even probability of predicting the wrong outcome if they do not change their forecast is worth the reputational gain from pooling with more competent experts. Relatively incompetent experts (those with $\pi \in \left[\frac{1}{2}, \pi^*\right)$) change their forecast. Their private information is sufficiently worse than the public information that the reputational gain from masquerading as a more competent expert is not worth the expected cost of an incorrect forecast.

The following intuition is useful in understanding the result. Clearly, an equilibrium cannot have $\pi^* > \pi^p$. If there were such an equilibrium, then any expert with competency $\pi \in (\pi^p, \pi^*)$ would find it preferable – both in terms of making the best forecast and in terms of his reputation – to defect from the equilibrium and not change his forecast. There also can be no equilibrium with $\pi^* = \pi^p$. If there were, all experts with π below, but sufficiently close, to π^p would defect. The expected increased cost of making the incorrect forecast would be negligible compared to the jump in their reputation.

A formal proof of the proposition is found in the appendix, but a sketch is as follows. By equation (9), an equilibrium with $\pi^* \in \left(\frac{1}{2}, 1\right)$ satisfies $\chi\left[\Pi^N(\pi^*) - \Pi^C(\pi^*)\right] = 1 - 2\theta(\pi^*, \pi^p)$. The right-hand side of this equation is the expert's expected cost if he does not change his forecast in the face of conflicting information and it is equal to the likelihood that he is correct if he does not change his forecast minus the likelihood that he is correct if he does change his forecast. It is decreasing in π^*, going to $\frac{1}{2}$ as π^* goes to $\frac{1}{2}$ and to zero as π^* goes to π^p. This is shown in Figure 1, drawn for $\pi^p = .75$. The left-hand side of the equation is the expert's reputational gain if he does not change his forecast. It is strictly positive and is demonstrated in the formal proof to be strictly increasing. The curve representing the left-hand side of the equation shifts up as χ increases and is shown in Figure 1 for $\chi = 1$. From the geometry, it is clear that as long as χ is not too large, the

are assumed to be mechanistic – also have no incentive to separate themselves.

curves representing the left- and right-hand sides cross exactly once at some $\pi^* \in \left(\frac{1}{2}, \pi^p\right)$. If χ is sufficiently large, the curve representing the right-hand side lies above the curve representing the left-hand side on $\left(\frac{1}{2}, 1\right)$ and the equilibrium has $\pi^* = \frac{1}{2}$.

An implication of Proposition 1 is that an increase in the quality of the public signal can increase the size of the set of experts who do not change their forecast when faced with conflicting and better quality public information. To see this, suppose that χ is sufficiently large that, given π^p, $\chi > \hat{\chi}$. Then experts pool: no expert changes his mind and the set of experts who continue to forecast an event that they believe is the less likely is $[\pi^*, \pi^p)$, where $\pi^* = \frac{1}{2}$. A marginal increase in π^p has no effect on π^* ($\hat{\chi}$ is continuous in π^p); hence, the set $[\pi^*, \pi^p)$ is enlarged.

Another – and striking – implication is that even when the public information is almost perfect, it is possible for all experts to continue to predict an event that they know is virtually certain *not* to occur. To see this, suppose that $\pi^p \to 1$. If $\pi^* = \frac{1}{2}$, then an expert who changes his forecast is believed to have competency $\frac{1}{2}$ and his revised forecast is correct with probability one. His payoff is thus $\frac{\chi}{2} + 1$. An expert who does not change his forecast is believed to have the average competency, conditional on initially forcasting incorrectly, of $\int_{1/2}^{1} p\,(1-p)\,dp / \int_{1/2}^{1} (1-p)\,dp = \frac{2}{3}$. His forecast is incorrect with probability one; hence his payoff is $\frac{2\chi}{3}$. As long as $\chi \geq 6$, it is an equilibrium for all experts to continue to forecast an event that will almost certainly not occur.

In this section, attention was restricted to equilibria where out-of-equilibrium beliefs were specified as the limits of equilibrium beliefs. However, other specifications of beliefs can result in other equilibria. In particular, there may be pooling equilibria where all experts change their forecast and an expert who changes his forecast is believed to have the average competency, conditional on initially forcasting incorrectly, of $\frac{2}{3}$. Such an equilibrium might be supported by the out-of-equilibrium belief that an expert who does not change his forecast is the *worst* possible type: his competency is $\frac{1}{2}$. To demonstrate that this is an equilibrium, it is sufficient to demonstrate that no expert would deviate from it. The expert with the most incentive to to deviate from it is the most competent

expert. Therefore, it is sufficient to show that an expert with $\pi = 1$ would not deviate. Such an expert would receive a payoff of $\frac{2\chi}{3}$ from following the equilibrium strategy and a payoff of $\frac{\chi}{2} + 1$ from deviating. Hence, if $\chi \geq 6$ such an equilibrium exists.

This type of pooling equilibrium is unappealing as the out-of-equilibrium beliefs are not sensible. Why would the market believe that an expert who deviates is the type of expert who has the *least* incentive to deviate? The problem, as previously noted, is that the perfect Bayesian equilibrium concept does not place any restrictions on out-of-equilibrium beliefs, other than that they support the equilibrium. It is typical to rule out pooling equilibria in signaling models that are supported by implausible beliefs by requiring that equilibria satisfy the D1 criterion.[8]

This is not the first paper to demonstrate that reputational or career concerns can distort decision making. The results in this section are related to the literature on anti-herding. In Avery and Chevalier (1999), two experts who care about being perceived as competent and who may have private information about their ability make forecasts in succession. If the second expert has sufficiently precise private information that he is of low competency, he may contradict the first expert's forecast with positive probability, even though he believes it likely that the first expert is correct. In Gilat (2004), experts who care about their reputations for competency and who have private information about their competency make a single forecast after observing public information. Experts of intermediate ability signal their competency by departing from the forecast favoured by the public information, even though their own information supports it. In these models, as in ours, anti-herding is a result of experts wanting to signal their private information. This contrasts with the herding that results in models where agents do not know their own private information.[9]

This section has demonstrated that optimising experts appear to ignore information that conflicts with their beliefs, even though they would make better decisions by con-

[8] See Ramey (1996). We discuss the D1 criterion in more detail in the next section.

[9] See, for example, Dewatripont, Jewitt and Tirole (1999).

sidering it. An interesting consequence of this is that more competent experts do not necessarily make better predictions. Although they make better first-round forecasts, by clinging to their original forecast in the face of superior contradictory evidence, experts with $\pi \in (\pi^*, \pi^p)$ make worse second-round forecasts than less competent experts.

3 Overconfidence

"Hurrah, boys, we've got them!" George Armstrong Custer at the Little Big Horn.[10]

"Major combat operations in Iraq have ended." George Bush, aboard the USS Abraham Lincoln, May 2003.

"The odds of a meltdown are one in 10,000 years." Ukrainian Minister of Power V. Sklyarov, February 1986.[11]

Overconfidence is pervasive.[12] A vast social psychology literature documents its existence. Most of us display it our own lives, in our certainty, for example, that we are better drivers than average.[13] Here we present a model where rational experts systematically exhibit overconfidence, even though there is no intrinsic benefit to doing so.

In this section we suppose that there is no publicly observed information, as there was in the previous section, but that after observing his own signal the expert announces the likelihood, ρ, that his forecast is correct. In terms of our banking story, ρ can be thought of as the vigour with which a banker attempts to sell his forecast to his employer and his clients. We assume that there is no direct benefit to an expert of announcing that he is correct with high probability and that there is a cost: if the expert turns out to be incorrect, he later suffers a loss that has a discounted present value of $c(\rho)$, where $c : \left[\frac{1}{2}, 1\right] \to \mathbb{R}_+$ is strictly increasing, concave, twice differentiable and has $c\left(\frac{1}{2}\right) = 0$.[14]

[10] Reported quote in Johnson (2004).

[11] Quoted in Gregorovich (1996).

[12] Overconfidence is sometimes called the "Lake Wobegone Effect" after the fictional Minnesota town where "all the children are above average".

[13] This is shown in numerous studies. Svenson (1981), for example, found that eighty percent of survey respondents claimed to be in the top thirty percent of all drivers.

[14] Concavity will turn out to be sufficient, but not necessary, for the second-order condition of the expert's problem to be satisfied.

We assume that the weight put on reputation is sufficiently high:

$$\chi \geq c'(1/2). \tag{10}$$

The payoff to the expert if he announces that he correct with probability ρ is then

$$\chi\Pi + P - c(\rho), \tag{11}$$

where χ, Π and P are defined as in the previous section.

We consider perfect Bayesian equilibria. As in the previous section, we need to make an assumption that rules out pooling equilibria based on implausible out-of-equilibrium beliefs.

Assumption 2. *Equilibria must satisfy the D1 criterion.*[15]

Intuitively, imposing the D1 criterion implies that following the observation of an off-the-equilibrium-path announcement, the public must put zero posterior weight on the expert being type π if there is another expert of type π' who has a greater incentive to deviate from the equilbrium, in the sense that type π' would strictly prefer to deviate for any resulting market belief Π that would make π weakly prefer deviating to not deviating.

If the D1 criterion holds, then the exquilibrium must be separating, except possibly for an interval of the most competent experts who claim to be right with probability one.

Proposition 2. *The equilibrium must have the following form: there is a* $\pi^* \in \left[\frac{1}{2}, 1\right]$ *such that experts in* $(\pi^*, 1]$ *say that they are certain that they are right. Experts with* $\pi \in \left[\frac{1}{2}, \pi^*\right)$ *separate: they each announce that they are correct with some probability in*

[15] Let $\{\rho(\pi), \Pi(\rho), \pi\}$ be a perfect Bayesian equilibrium. Let $\hat{\rho}$ be an out-of-equilibrium action and suppose that $\hat{\Pi} \in \left[\frac{1}{2}, 1\right]$ is the market's assessment of the expert's type if it observes such an action. Suppose that there is a non-emply set of expert types $S' \subset \left[\frac{1}{2}, 1\right]$ such that for every expert of type $\pi \notin S'$ who weakly prefers following the out-of-equilibrium strategy $\hat{\rho}$ and being thought to be type $\hat{\Pi}$ to following his equilibrium strategy $\rho(\pi)$ and being thought to be type $\Pi(\rho(\pi))$ there exists an expert of type $\pi' \in S'$ who strictly prefers following the out-of-equilibrium strategy $\hat{\rho}$ and being thought to be type $\hat{\Pi}$ to following his equilibrium strategy $\rho(\pi')$ and being thought to be type $\Pi(\rho(\pi'))$. Then the equilibrium violates the D1 criterion unless, upon observing $\hat{\rho}$, the market infers that the expert's type $\pi \in S'$.

$\left[\frac{1}{2}, 1\right)$ *and their announcement reveals their type.*

Note that the proposition does not rule out $\pi^* = \frac{1}{2}$ or $\pi^* = 1$. The formal proof, in the appendix, borrows from Ramey (1996). The strategy of the proof is to demonstrate that if the D1 criterion holds and if any two experts of different types pool at any announcement other than $\rho = 1$, then the more competent expert has an incentive to deviate. Pooling with $\rho = 1$ is not ruled out by the D1 criterion, but there is no equilibrium where a less competent expert chooses $\rho = 1$ and a more competent expert chooses $\rho < 1$. Suppose that the market believes that an expert who chooses $\rho = 1$ is more competent than an expert who chooses $\rho < 1$. Then if the less competent expert is willing to choose $\rho = 1$ to be thought more competent, then the more competent expert must be willing as well.

The market conjectures that a policy maker of type $\pi \in \left[\frac{1}{2}, \pi^*\right)$ announces that he is correct with probability $\rho_c(\pi) < 1$. Separability implies that $\rho_c : \left[\frac{1}{2}, \pi^*\right] \to \left[\frac{1}{2}, 1\right)$ is one to one. Hence, upon observing $\rho < 1$, the market infers that the expert is type $\rho_c^{-1}(\rho)$. Thus

$$\Pi = \begin{cases} \frac{\pi^*+1}{2} \text{ if } \rho = 1 \\ \rho_c^{-1}(\rho) \text{ otherwise.} \end{cases} \tag{12}$$

In equilibrium, the market's conjecture must be correct and $\rho_c(\pi^*) = \rho(\pi^*)$. Suppose that there is an interior threshold $\pi^* \in \left(\frac{1}{2}, 1\right)$. Then the threshold expert – the one with competency $\pi = \pi^*$– must be indifferent between announcing that he is correct with probability one and announcing that he is correct with probability $\rho(\pi^*)$. If he claims to be correct with probability one, then by equation (12), the market's assessment of his competency is $\frac{\pi^*+1}{2}$. If his forecast turns out to be incorrect, then he suffers a loss of $c(1)$. Thus, by equation (11), his expected payoff is $\chi\frac{\pi^*+1}{2} + \pi^* - (1-\pi^*)c(1)$. If, instead, he claims to be correct with probability $\rho(\pi^*)$, then he is thought to have competency π^*. If his forecast turns out to be incorrect, then he incurs a loss of $c(\rho(\pi^*))$. Thus, by equation (11), his expected payoff is $\chi\pi^* + \pi^* - (1-\pi^*)c(\rho(\pi^*))$. Equating the expected payoff from claiming to be correct with probability one to the expected payoff from claiming to

be correct with probability $\rho(\pi^*)$ yields

$$\pi^* = \begin{cases} \rho^{-1}\left(c^{-1}\left(c(1) - \frac{\chi}{2}\right)\right) \in \left(\frac{1}{2}, 1\right) \text{ if } \chi < 2c(1) \\ \frac{1}{2} \text{ otherwise.} \end{cases} \tag{13}$$

If $\pi^* \in \left(\frac{1}{2}, 1\right]$ and $\pi < \pi^*$, then by equations (11) and (12), the expert maximises

$$\chi \rho_c^{-1}(\rho) - (1-\pi)\, c(\rho). \tag{14}$$

We conjecture that $\rho_c(\pi)$ is twice differentiable and it will later be clear that this is the case.[16] The first- and second-order conditions for a solution to the expert's problem are:

$$\chi \rho_c^{-1\prime}(\rho) - (1-\pi)\, c'(\rho) = 0 \tag{15}$$

$$\chi \rho_c^{-1\prime\prime}(\rho) - (1-\pi)\, c''(\rho) = 0 \tag{16}$$

Using the rules $f^{-1\prime}(x) = 1/f'(x)$ and $f^{-1\prime\prime}(x) = -f''(x)/f'(x)^3$ and imposing $\rho_c(\pi) = \rho(\pi)$, equations (15) and (16) yield

$$\frac{\chi}{(1-\pi)\, c'(\rho(\pi))} = \rho'(\pi) \tag{17}$$

$$-\frac{\chi \rho''(\pi)}{\rho'(\pi)^2} - (1-\pi)\, c''(\rho(\pi)) < 0. \tag{18}$$

Equation (17) is a first-order condition with no boundary condition. Following Riley (1979), it is conventional in signalling models to generate a boundary condition by assuming that the agent with the lowest-quality private information (here, an expert with $\pi = \frac{1}{2}$) would not send a costly signal. The logic is that, in a separating equilibrium, expectations can be no worse; hence there is no point to costly signalling.[17] Thus

$$\rho(1/2) = 1/2. \tag{19}$$

[16] There are no separating equililbria that are not differentiable. See Mailath (1987).

[17] This is also the only equilibrium that satisfies the D1 criterion.

Definition 2. *An equilibrium is a* $\pi^* \in \left[\frac{1}{2}, 1\right]$ *and a twice-differentiable function* $\rho(\pi) : \left[\frac{1}{2}, \pi^*\right] \to \left[\frac{1}{2}, 1\right]$ *such that equations (13) and (17) - (19) are satisfied.*

An equilibrium is a perfect Bayesian equilibrium. The expert is maximising his payoff while taking into account the effect of his action on the beliefs of the market. The market's beliefs are (trivially) consistent with Bayes rule and are formed using the correct conjecture about the equilibrium strategies and the observation of ρ.

Proposition 3. *If* $\chi \geq 2c(1)$*, then all experts claim that they are correct with probability one. Furthermore, if* $\chi < 2c(1)$*, then all experts with competencies in* $[\pi^*, 1]$ *claim that they are the correct with probability one and experts with* $\pi \in \left[\frac{1}{2}, \pi^*\right)$ *claim that they are correct with probability* $\rho(\pi)$*, where*

$$\rho(\pi) = c^{-1}\left(-\chi \ln\left(2\left(1-\pi\right)\right)\right) \text{ and } \rho(\pi) > \pi, \ \pi \in \left[\frac{1}{2}, \pi^*\right) \tag{20}$$

$$\pi^* = 1 - \frac{1}{2}\exp\left(\frac{1}{2} - \frac{c(1)}{\chi}\right) \in \left(\frac{1}{2}, 1\right). \tag{21}$$

Proof. See the Appendix.

Proposition 2 demonstrates that experts with competencies in the separating region $\left[\frac{1}{2}, \pi^*\right)$ are overconfident, as well as experts in the pooling region $[\pi^*, 1]$. It also ensures that $\pi^* < 1$ and equilibria with complete separation do not exist. The intuition is that, in the separating region, experts with good quality private information separate themselves from senders of poorer quality information by saying that they are more confident than they actually are. However, there is an upperbound on how overconfident an expert can be: ρ can be no greater than one. Thus, experts with very good quality information are unable to separate themselves.[18]

The model here predicts the overconfidence documented in the social psychology literature. It also predicts a number of properties of this overconfidence. First, overconfidence occurs when people are rewarded based on the perceived abilities. It does not exist if

[18] Cho and Sobel (1990) consider a general game where the sender of the signal's action space is bounded above and find that a possible outcome is a set of types pooling at the highest possible action.

people are rewarded solely on their performance (that is, when $\chi = 0$). Second, it is possible for experts with a sizable range of competencies to insist that they are certain that they are right. Third, confidence is (weakly) increasing in competency and if χ is sufficiently small, then experts of intermediate competency display the most overconfidence: the polar experts with competencies $\pi = \frac{1}{2}$ and $\pi = 1$ do not overstate their competency. If χ is large enough, then it is the least competent experts who are the most overconfident. It is interesting to ask how these three predictions of our model match the empirical evidence.

We find some evidence that is consistent with the first property. While overconfidence is widespread, a few types of experts appear to exhibit little or no overconfidence. Examples are bridge players, oddsmakers and weather forecasters.[19] For all of these people, the success or failure of their conjectures is immediately and publicly observable. Hence, it is likely that they perceive their reward to be based on their performance rather than their perceived competency. There is significant evidence in favour of the second property. Fischhoff et al (1977) found that when people claimed to be 100 percent confident, they were right about 70 - 80 percent of the time.

The third property is less consistent with the evidence, however. In particular, the evidence suggests that not only do the least competent subjects overstate their competency the most, as we predict when χ is sufficiently high, but they are also the ones who claim to be the most confident. The failure of our model to predict this may be due to our assumption that people have perfect knowledge of their own competency. Kruger and Dunning (1999) argue that incompetency robs people of the ability to realise that they are inaccurate. This, they suggest, is responsible for data showing that incompetent people dramatically overestimate their abilities relative to more competent people.

We are not the first to explain overconfidence in an optimising model; alternative frameworks are offered by Van den Steen (2004) and Brocas and Carrillo (2002). Van den Steen (2004) explains overconfidence by supposing that individuals have noisy idio-

[19] See Plous (1993) for a survey of this literature.

syncratic information. Thus, agents who select an option are more likely to be optimistic about their choice than other agents. Brocas and Carrillo (2002) suppose that agents choose between a riskless activity and an activity that can yield either a high or a low payoff, depending upon their competency. If agents are uncertain about their abilities and information acquisition is costly they choose the risky activity if preliminary evidence about their competency is positive; they do not choose the safe activity without substantial information that they are incompetent. Thus, it is more likely that incompetent people will engage in the risky activity than it is that competent people will engage in the safe activity.

While we have demonstrated that overconfidence is consistent with optimising behaviour, it is so ubiquitous a phenomenon that it must have other causes than just the one that we suggest. Suppose that some of the bankers' overconfidence was the result of cognitive errors that are not explained by the model. How might the overconfidence have been reduced? Koriat et al (1980) found that getting subjects to search for reasons that disconfirmed their hypotheses reduced overconfidence. In the next section, we present a model where bankers have an incentive to refuse to do this.

3.1 Disconfirming Evidence

"It is the peculiar and perpetual error of human intellect to be more moved and excited by affirmatives than by negatives." Francis Bacon, 1620.[20]

In a famous early study, Wason (1960) presented subjects with a triple, (2,4,6) and told them that the triple conformed to a particular rule. The subjects were asked to find the rule by generating their own triples and presenting them for positive or negative feedback. He found that subjects had great difficulty finding the rule, which was "any ascending sequence". The problem was that the subjects appeared to test only triples that conformed to their rule and not those that would have disconfirmed their theory. According to psychologists Oaksford and Chater (1993), this and subsequent similar

[20]Reported in Plous (1993).

studies have "raised more doubts over human rationality than any other psychological tasks."

We consider a scenario where an expert must predict which one of three or more events will occur. He receives a signal that tells him that one of the events will occur with probability $\pi \in \left[\frac{1}{2}, 1\right]$. As in the previous sections, this likelihood that his signal is correct is referred to as his *competency* and it is his private information. After receiving the signal the expert forecasts the event that his signal favours. He then has the opportunity to invest in the possibility of finding disconfirming information. Specifically, if the expert pays a cost q, then if his signal was incorrect he receives private information that it is incorrect with probability π_d. The expert can then continue with his original prediction or he can withdraw his forecast.[21] Later, the event is observed and it is learned if the expert was correct or not.

The expert's payoff is

$$\chi\Pi + P^d - \delta q, \tag{22}$$

where the variable P^d equals one if he persists with his original forecast and it is correct, zero if he withdraws his original forecast and minus one if he persists in his original forecast and it turns out to be incorrect. The variable δ equals one if he invests in additional information and zero otherwise. The parameter χ and the variable Π are as defined in the previous sections.

As a benchmark, we first consider the case where $\chi = 0$. Suppose that an expert invests in the possibility of finding disconfirming evidence. With probability π, his original forecast is correct and he finds no disconfirming information. Thus, he continues with his original choice, which he knows to be correct with probability greater than one half, and is later proved to be correct. Hence, $P^d = 1$. With probability $(1 - \pi)\,\pi_d$, his original forecast is wrong and he receives confirmation of this. He withdraws his original forecast

[21] Because there are more than two possible events, disconfirming evidence does not resolve the uncertainty. Upon receiving proof that his original forecast was wrong, the expert no longer has information that is useful to the market. Thus, changing his forecast is not an option in this scenario.

and $P^d = 0$. With probability $(1-\pi)(1-\pi_d)$ his original forecast is wrong, but he does not receive disconfirming information. He continues with his original choice, which he believes is correct with probability greater than one half, and $P^d = -1$. Thus, the expected value of P^d is $\pi - (1-\pi)(1-\pi_d)$ and the expert's payoff when he does not invest in the possibility of finding disconfirming evidence is $\pi - (1-\pi)(1-\pi_d) - q$.

If the expert does not invest in the possibility of finding disconfirming evidence, then he continues with his original forecast and his payoff is equal to the expected value of P^d, which is equal to the probability his choice is correct minus the probability it is not, or $\pi - (1-\pi)$. The expert will choose to invest in the possibility of receiving disconfirming information if the expected payoff from doing so exceeds the expected payoff from not doing so. This is the case when

$$1 - \frac{q}{\pi_d} =: \tilde{\pi} > \pi. \tag{23}$$

Thus, if there are no reputational considerations, it is the *less* competent experts who invest in acquiring disconfirming information; relatively competent experts do not. This is because, as their own signal is more likely to be correct, relatively competent experts find that a search for evidence proving otherwise is less likely to be informative.[22] We assume that the cost of acquiring information is sufficiently low that, in the absence of reputational concerns, some experts would acquire it: $\pi_d > 2q$.

We now suppose that reputational concerns matter, that is $\chi > 0$, and we initially suppose that the investment in information is observable, although the result is not. We look for a threshold equilibrium where experts with $\pi \in \left[\frac{1}{2}, \pi^*\right)$ invest in information acquisition and experts with $\pi \in [\pi^*, 1]$ do not.

Let Π^D be the expert's reputation if he does not invest in additional information, Π^N be his reputation if he does invest and does not withdraw his original forecast and Π^W be

[22] The argument that, even without reputational concerns, competent experts are unlikely to look for disconfirming evidence because they are unlikely to find it is related to Oaksford and Chater's (1994) argument that a failure to focus solely on evidence that might disprove a hyphothesis may be a result of the properties that figure in a causal relationship being rare.

his reputation if he invests and then withdraws his forecast. We consider equilibria where $\Pi^D > \Pi^N > \Pi^W$.[23] In such equilibria an expert who invests in additional information and does not receive disconfirming evidence continues with his original forecast: this is best both in terms of maximising his payoff from making the best forecast and enhancing his reputation. An expert who invests in additional information and receives disconfirming evidence withdraws his forecast. This is because an equilibrium strategy of investing in additional information and persisting with his original forecast with strictly positive probability in the face of disconfirming evidence must have the same payoff as the strategy of investing in additional information and always persisting with his original forecast. This latter strategy is dominated by the strategy of not investing in additional information.

The payoff to the strategy of not investing in additional information is $\chi\Pi^D + \pi - (1-\pi)$. The payoff to the strategy of investing in additional information and withdrawing one's forecast if and only if disconfirming evidence is found is $[1-\pi(1-\pi_d)]\Pi^N + (1-\pi)\pi_d\Pi^W + \pi - (1-\pi)(1-\pi_d) - q$. By the same reasoning as in section 1, the expert follows a threshhold strategy. There is a π^* such that the expert invests in additional information if $\pi < \pi^*$, does not invest if $\pi > \pi^*$ and is indifferent over randomisations if $\pi = \pi^*$. Thus, in equilibrium the market conjectures that the expert follows such a threshhold strategy. Thus, using Bayes' rule (as in equation (1)), we have

$$\Pi^A = \begin{cases} \Pi^D(\pi^*) = \frac{1+\pi^*}{2} \text{ if } A = D \text{ and } \pi^* < 1 \\ \Pi^N(\pi^*) = \frac{\int_{1/2}^{\pi^*} p[1-\pi_d(1-p)]dp}{\int_{1/2}^{\pi^*}[1-\pi_d(1-p)]dp} \text{ if } A = N \text{ and } \pi^* > \frac{1}{2} \\ \Pi^W(\pi^*) = \frac{\int_{1/2}^{\pi^*} p(1-p)dp}{\int_{1/2}^{\pi^*}(1-p)dp} \text{ if } A = W \text{ and } \pi^* > \frac{1}{2}. \end{cases} \tag{24}$$

We specify out-of-equilibrium beliefs as in Assumption 1.

Assumption 3. *$\Pi^D(1) = \lim_{\pi^*\to 1}\Pi^D(\pi^*) = 1$ and $\Pi^A\left(\frac{1}{2}\right) = \lim_{\pi^*\to 1/2}\Pi^A(\pi^*) = \frac{1}{2}$, $A = N, W$.*

We have the following definition.

[23] As in the previous sections, there may be equilibria supported by unappealing out-of-equilibrium beliefs that do not satisfy this condition.

Definition 3. *An **equilibrium** is a π^*such that*

$$\pi^* \left\{ \begin{array}{c} = \frac{1}{2} \\ \in [0,1] \\ = 1 \end{array} \right\} \text{ and } \chi \Pi^D(\pi^*) + q \left\{ \begin{array}{c} \geq \\ = \\ \leq \end{array} \right\}$$
$$[1-(1-\pi^*)\pi_d]\,\Pi^N(\pi^*) + (1-\pi^*)\pi_d \Pi^W(\pi^*) + (1-\pi^*)\pi_d \tag{25}$$

Proposition 4. *If $\chi \geq 2(\pi_d - 2q)$ then there is a unique equilibrium where no expert invests in the possibility of finding disconfirming evidence. Furthermore, if $\chi < 2(\pi_d - 2q)$, then there is a unique $\pi^* < \tilde{\pi}$ such that experts with competency $\pi < \pi^*$ invest in the possibility of finding disconfirming evidence and experts with competency $\pi > \pi^*$ do not.*

Thus, we have that when the search for disconfirming evidence is observable, fewer experts will invest in the possibility of finding disconfirming evidence than they would if they did not have reputational concerns. If the reputational concerns are important enough, no expert will invest in the possibility of finding disconfirming evidence.[24]

We now consider the case where an investment in the possibility of finding disconfirming evidence is unobservable. In this case, the market observes only whether the expert continues to maintain his original forecast (action N) or withdraws it (action W). If an expert does not withdraw his forecast then the market believes that either the expert did not invest in the possibility of finding disconfirming information and, hence, $\pi \in [\pi^*, 1]$ or that the expert did invest, and hence $\pi \in \left[\frac{1}{2}, \pi^*\right]$, but no disconfirming evidence was received. Thus, if an expert does not withdraw his original forecast, the market's assessment of his competency is

$$\Pi^A = \left\{ \begin{array}{l} \Pi^N(\pi^*) = \frac{\int_{1/2}^{1} p dp - \pi_d \int_{1/2}^{\pi^*} p(1-p)dp}{\int_{1/2}^{1} dp - \pi_d \int_{1/2}^{\pi^*} (1-p)dp} \text{ if } A = N \\ \Pi^W(\pi^*) = \frac{\int_{1/2}^{\pi^*} p(1-p)dp}{\int_{1/2}^{\pi^*} (1-p)dp} \text{ if } A = W \text{ and } \pi^* > \frac{1}{2}. \end{array} \right. \tag{26}$$

[24] The case of confirming evidence is not symmetric. In the absence of reputational concerns, especially competent agents would not engage in a costly search for confirming evidence. If they did, they would likely find it, but it would not be particularly useful: they already know they are likely correct. Relatively incompetent experts would not search either: it is too unlikely they would receive information.

As before, we have:

Assumption 4. $\Pi^W\left(\frac{1}{2}\right) = \lim_{\pi^* \to 1/2} \Pi^W(\pi^*) = \frac{1}{2}$.

If an expert invests in the possibility of finding disconfirming information, then his expected payoff is $\chi\left\{\left[\pi + (1-\pi)(1-\pi_d)\right]\Pi^N(\pi^*) + (1-\pi)\pi_d\Pi^W(\pi^*)\right\} + \pi - (1-\pi)(1-\pi_d)$ - q. If he does not invest in the possibility of finding disconfirming information, then his expected payoff is $\chi\Pi^N(\pi^*) + \pi - (1-\pi)$. The threshold expert is indifferent; hence, I have the following.

Definition 4. *An **equilibrium** is a π^*such that*

$$\pi^* \left\{ \begin{array}{c} = \frac{1}{2} \\ \in [0,1] \\ = 1 \end{array} \right\} \text{ and } (1-\pi^*)\pi_d - q \left\{ \begin{array}{c} \geq \\ = \\ \leq \end{array} \right\} \chi(1-\pi^*)\pi_d\left[\Pi^N(\pi^*) - \Pi^W(\pi^*)\right]. \quad (27)$$

Proposition 5. *If $\chi \geq 4(\pi_d - 2q)/\pi_d$ then there is an equilibrium where no expert invests in the possibility of finding disconfirming evidence. Furthermore, if $\chi < 4(\pi_d - 2q)/\pi_d$, then there is a $\pi^* < \tilde{\pi}$ such that experts with competency $\pi \leq \pi^*$ invest in the possibility of finding disconfirming evidence and experts with competency $\pi \leq \pi^*$ do not.*

In this and the previous two sections, the desire of experts to be seen as competent distorts their predictions. Does this imply that their behaviour is harmful? Is what looks like confirmation bias necessarily a bad thing? As is typical in signalling models, the experts here engage in costly behaviour to manipulate the market's beliefs: they distort their forecasts and this is damaging. Less typically, perhaps, the signalling does not necessarily convey more information about the expert's competency that might be beneficial to the market. In the belief persistence model of section one and in the model of this section, in the absence of reputational concerns, the experts would split into two pools, one consisting of more competent experts and one consisting of less competent experts. With relatively weak reputational concerns, the experts still split into two pools,

although there would be more experts in the relatively competent pool and fewer in the less competent pool. With strong reputational concerns, however, there is complete pooling and the market has *less* information than it would have without reputational concerns. In the overconfidence model of section two, in the absence of reputational concerns the experts would pool: no one would exhibit overconfidence. With strong reputational concerns there would also be pooling: everyone would exhibit the same perfect confidence. However, with weak reputational concerns there is some separation and the market does gain some information relative to what it would have with no reputational concerns.

While we have provided an explanation for not considering evidence that would disconfirm one's hypotheses that is consistent with rational optimising behaviour, there are undoubtedly other explanations as well. Westen at al (2006) provide a physiological one. They used neuroimaging to study the brains of party loyalists during the 2004 US Presidential election. Subjects were confronted with reasoning tasks involving information damaging to their candidate, the other candidate or some neutral control target. They found that when subjects had an emotional stake, there was neural activity in different parts of the brain than when they did not. This supports a belief that the brain seeks solutions that satisfy emotional, as well as cognitive, constraints.

4 Conclusion

In this paper, we demonstrate that a desire to appear competent may explain what appear to be the cognitive errors that are known as confirmation bias. Our motivation is an explanation of the behaviour of bankers and other employees of the financial services industry – people whose rewards are determined by their perceived ability, as well as their long-term performance. However, the model of this paper may explain behaviour resembling confirmation bias in other scenarios as well. Examples include policy makers who persist in policies long after there is substantial evidence that the policies are not in

their best interests (see Tuchman (19984) and scientists who stubbornly refuse to accept theories contradicting their own (see Nickerson (1998)).

In addition to contributing to the recent excesses in financial markets, it is possible that a reward structure that leads to confirmation bias can also cause asset price anomalies. While not providing reasons for the phenomenon, a sizable behavioural finance literature attempts to explain empirical pricing puzzles as the result of confirmation bias, in particular, overconfidence. Daniel et al (1998), for example, assume that market participants are overconfident, in the sense of believing that their information is more accurate than that of the market, and they demonstrate that overconfidence implies the negative long-lag autocorrelations and excess volatility found in stock market data.[25]

Appendix

Proof of Proposition 1. Define $L(\pi^*, \pi^p) := \chi\left[\Pi^N(\pi^*, \pi^p) - \Pi^C(\pi^*, \pi^p)\right]$ and $R(\pi^*, \pi^p) := 1 - 2\theta(\pi^*, \pi^p)$.The proposition follows from the following properties of L and R: (i) $\partial L/\partial \pi^* > 0$; (ii) $L\left(\frac{1}{2}, \pi^p\right) = \frac{\chi(5-4\pi^p)}{6(3-2\pi^p)} > 0$; (iii) $\partial R/\partial \pi^* < 0$; (iv) $R\left(\frac{1}{2}, \pi^p\right) = 2\pi^p - 1$; (v) $R(\pi^p, \pi^p) = 0$. Properties (ii) - (v) are straightforward. We show property (i). By equations (2) and (8)

$$\Pi^A(\pi^*, \pi^p) = \begin{cases} \frac{2\pi^*}{3} + \frac{1}{3(2\pi^p-1)}\left[\pi^p - \frac{2(1-\pi^p)^2}{1-(2\pi^p-1)\pi^*}\right] & \text{if } A = N \\ \frac{2\pi^*}{3} + \frac{1}{3(2\pi^p-1)}\left[\pi^p - \frac{\frac{1}{2}}{\frac{1}{2}+\pi^p-(2\pi^p-1)\pi^*}\right] & \text{if } A = C. \end{cases} \tag{A1}$$

By equation (A1), $\partial L/\partial \pi^* > 0$ iff $1/\left[\frac{1}{2} + \pi^p - (2\pi^p - 1)\pi^*\right]^2 > 4(1-\pi^p)^2/\left[1-(2\pi^p-1)\pi^*\right]^2$. This is true iff $\left[1-(2\pi^p-1)\pi^*\right] > 2(1-\pi^p)\left[\frac{1}{2} + \pi^p - (2\pi^p-1)\pi^*\right]$.As both sides are linear in π^p this is true if it is true at the endpoints $\pi^p = \frac{1}{2}$ and $\pi^p = 1$. This is straightforward to show.

Proof of Proposition 2. Let $\rho(\pi)$ be the equilibrium strategy of a type-π expert and let $\Pi(\rho)$ be the market's equilibrium assessment of π given an observation of ρ. We first demonstrate that there cannot be pooling at any $\rho < 1$. Suppose to the contrary that

[25] See Glaser (2004) for a survey. More generally, Hirschleifer (2001) provides a survey of different types of departures from perfect rationality and evidence of their existence in financial markets.

there exists a pooling equilibrium with $\rho(\pi) = \rho'$ for more than one $\pi \in \left[\frac{1}{2}, 1\right]$. Let $\pi' = \sup\{\pi | \rho(\pi) = \rho'\}$. Clearly $\Pi(\rho') < \pi'$. Choose $\pi'' \in (\Pi(\rho'), \pi')$.

By equation (11), if $\rho = \rho'$, then a marginal increase in ρ accompanied by an increase in Π raises the welfare of a type-π expert if and only if $d\Pi > \frac{1}{\chi}(1-\pi) c'(\rho') d\rho'$. Thus, since $\pi'' < \pi'$, for $\tilde{\rho}$ strictly greater than, but sufficiently close to, ρ' it is possible to find a $\tilde{\Pi} > \Pi(\rho')$ such that the type-π' expert strictly prefers choosing $\tilde{\rho}$ and being thought type $\tilde{\Pi}$ to choosing ρ' and being thought type $\Pi(\rho')$ and a type-π expert strictly prefers choosing ρ' and being thought type $\Pi(\rho')$ to choosing $\tilde{\rho}$ and being thought type $\tilde{\Pi}$, for every $\pi \leq \pi''$. Choose $\tilde{\rho}$ sufficiently close to ρ' that $\tilde{\Pi} < \pi''$.

Case 1. Suppose that $\tilde{\rho}$ is an out-of-equilibrium action. If type $\pi \leq \pi''$ weakly prefers prefers choosing $\tilde{\rho}$ and being thought type Π to choosing ρ' and being thought type $\Pi(\rho')$ then it must be that $\Pi > \tilde{\Pi}$. We also have that type π' strictly prefers choosing $\tilde{\rho}$ and being thought type Π to choosing ρ' and being thought type $\Pi(\rho')$. Thus, for the D1 criterion to hold it must be that if the public observes the out-of-equilibrium action $\tilde{\rho}$ it believes that $\pi > \pi''$. Thus, $\Pi(\tilde{\rho}) > \pi''$. Since type π' strictly prefers choosing $\tilde{\rho}$ and being thought type $\tilde{\Pi}$ to choosing ρ' and being thought type $\Pi(\rho')$, he strictly prefers choosing $\tilde{\rho}$ and being thought type $\Pi(\tilde{\rho}) > \tilde{\Pi}$ to choosing ρ' and being thought type $\Pi(\rho')$. Thus, type π' defects. This is a contradiction.

Case 2. Suppose that $\tilde{\rho}$ is an equilibrium action. We have that the type-π' expert strictly prefers choosing $\tilde{\rho}$ and being thought type $\tilde{\Pi}$ to choosing ρ' and being thought type $\Pi(\rho')$; hence, $\tilde{\Pi} > \Pi(\tilde{\rho})$. We have that a type-$\pi$ expert strictly prefers choosing ρ' and being thought type $\Pi(\rho')$ to choosing $\tilde{\rho}$ and being thought type $\tilde{\Pi}$, for every $\pi \leq \pi''$. Thus, no expert of type-π, $\pi \leq \pi''$, chooses $\tilde{\rho}$. This implies that $\Pi(\tilde{\rho}) \geq \pi'' > \tilde{\Pi}$: a contradiction.

This concludes the first part of the proof. We now show that if an equilibrium has a type-π' expert choosing $\rho = 1$ then each type-π expert, $\pi > \pi'$ also chooses $\rho = 1$. Suppose to the contrary that an equilibrium has a type-π' expert choosing $\rho = 1$ and a type-π, $\pi > \pi'$, expert choosing $\tilde{\rho} < 1$. By equation (11), for the experts not to defect we

require $1-\pi' \leq \chi\left[\Pi\left(1\right)-\Pi\left(\tilde{\rho}\right)\right]/\left[c\left(1\right)-c\left(\tilde{\rho}\right)\right] \leq 1-\pi$: a contradiction.

Proof of Proposition 3. The differential equation (17) is separable and has solutions $\rho\left(\pi\right)=c^{-1}\left(k-\chi\ln\left(1-\pi\right)\right)$, where k is a constant. Imposing the boundary condition (20) yields equation (20). Solving equations (13) and (20) yields equation (21). Differentiating equation (17) yields

$$-\frac{\rho''\left(\pi\right)}{\rho'\left(\pi\right)^{2}}=\left(1-\pi\right)c''\left(\rho\left(\pi\right)\right)\rho'\left(\pi\right)-c'\left(\rho\left(\pi\right)\right). \tag{A2}$$

Assumption (10) and the concavity of c ensures that $\rho'\left(\pi\right)>1$. This and equation (A2) ensure that condition (18) is satisfied. Equation (20) and $\rho'\left(\pi\right)>1$ $\rho\left(\pi\right)>\pi$ for $\pi\in\left(\frac{1}{2},\pi^{*}\right]$.

Proof of Proposition 4. By Assumption 3 and equations (24) and (25), $\pi^{*}=\frac{1}{2}$ iff $\chi\geq 2\left(\pi_{d}-2q\right)$ and $\pi^{*}\neq 1$. Suppose $\chi<2\left(\pi_{d}-2q\right)$ and let $R_{0}\left(\pi^{*}\right):=\Pi^{D}\left(\pi^{*}\right)-\Pi^{N}\left(\pi^{*}\right)$ and $R_{1}\left(\pi^{*}\right):=\pi_{d}\left(1-\pi^{*}\right)\left[\Pi^{N}\left(\pi^{*}\right)-\Pi^{W}\left(\pi^{*}\right)\right]$.Then by equation (24)

$$R_{0}\left(\pi^{*}\right) = \frac{3-\pi^{*}}{6}-\frac{1}{3}\frac{\frac{3}{2}-\pi^{*}-\pi_{d}\left(1-\pi^{*}\right)}{2-\pi_{d}\left(\frac{3}{2}-\pi^{*}\right)} \tag{A3}$$

$$R_{1}\left(\pi^{*}\right) = \frac{\pi_{d}\left(1-\pi^{*}\right)\left(\pi^{*}-\frac{1}{2}\right)^{2}}{3\left(\frac{3}{2}-\pi^{*}\right)\left[2-\pi_{d}\left(\frac{3}{2}-\pi^{*}\right)\right]}. \tag{A4}$$

By equation (25), an interior threshold requires

$$\left(1-\pi^{*}\right)\pi_{d}-q=\chi\left[R_{0}\left(\pi^{*}\right)+R_{1}\left(\pi^{*}\right)\right]. \tag{A5}$$

The left-hand side of equation (A5) is strictly decreasing, equaling $\frac{\pi_{d}}{2}-q$ when $\pi^{*}=\frac{1}{2}$ and equaling zero as $\pi^{*}=\tilde{\pi}$. The right-hand side of equation (A5) is strictly positive, equalling $\frac{\chi}{4}<\frac{\pi_{d}}{2}-q$ when $\pi^{*}=\frac{1}{2}$ and equalling $\frac{\chi}{3}\frac{3-\pi_{d}}{4-\pi_{d}}>0$ when $\pi^{*}=1$. Thus, $\exists\pi^{*}\in\left[\frac{1}{2},\tilde{\pi}\right]$ such that equation (A5) holds.

This π^{*} is unique if the right-hand side is increasing, or if it is decreasing and less steep than the left-hand side. This is the case if $\pi_{d}>\chi\left[R_{0}'\left(\pi^{*}\right)+R_{1}'\left(\pi^{*}\right)\right]$. We have

$\frac{\chi}{2} < \pi_d - 2q < \pi_d$; hence this is true if

$$\frac{1}{2} + R_0'(\pi^*) + R_1'(\pi^*) > 0. \tag{A6}$$

By equations (A3) and (A4) We have

$$R_0'(\pi^*) = \frac{\pi_d}{6} \frac{2 + \pi_d - 4\pi^* - \pi_d x^2}{(2 - \pi_d x)^2} \tag{A7}$$

$$R_1'(\pi^*) = -\frac{\pi_d}{6} \frac{x(1-x)(2-\pi_d x) - 2(2x-1)x^2(1-\pi_d x)}{x^2(2-\pi_d x)^2} \tag{A8}$$

where $x \equiv \frac{3}{2} - \pi^* \in \left(\frac{1}{2}, 1\right)$. Substituting equations (A7) and (A8) into inequality (A6) yields

$$\begin{aligned} &3x^2(2-\pi_d x)^2 + \pi_d\left(\pi_d - 4 + 4x - \pi_d x^2\right) \\ &-\pi_d x(1-x)(2-\pi_d x) + 2\pi_d(2x-1)(1-x)^2(1-\pi_d x) > 0 \end{aligned} \tag{A9}$$

The left-hand side of inequality (A9) is decreasing in π_d if and only if

$$-8x^3 + 2\pi_d - 6 + 10x - 8x^2 - 16\pi_d x^2 + 18\pi_d x^3 - 2\pi_d x^4 + 4\pi_d x < 0. \tag{A10}$$

The left-hand side of inequality (A10) is linear in π_d and, hence, it can be verified to hold by checking the endpoints. We have $-8x^3 - 6 + 10x - 8x^2 < 0$ when $\pi_d = 0$ and $-4 + 14x - 24x^2 + 10x^3 - 2x^4 < 0$ when $\pi_d = 1$. Thus, it is sufficient to show that inequality (A10) holds at $\pi_d = 1$. This is true if and only if $G(x) := -x^4 + x^3 - 4x^2 + 12x - 5 > 0$. An interior minimum for G requires $F := -4x^3 + 3x^2 - 8x + 12 = 0$. However, F has no roots in $\left(\frac{1}{2}, 1\right)$; hence it is sufficient to show that $G > 0$ at the endpoints. I have $G\left(\frac{1}{2}\right) = \frac{1}{16} > 0$ and $G(1) = 3 > 0$.

Proof of Proposition 5. We have that $(1 - \pi^*)\pi_d - q$ is strictly decreasing in π^*, equalling $\pi_d/2 - q$ when $\pi^* = \frac{1}{2}$ and equalling zero when $\pi^* = \tilde{\pi}$. By equations (26)

and (27), $\chi(1-\pi^*)\pi_d\left[\Pi^N(\pi^*)-\Pi^W(\pi^*)\right]$ is strictly positive and equals $\chi\pi_d/8$ when $\pi^*=\frac{1}{2}$. This yields the result.

References

Brocas, Isabelle and Juan D. Carrillo, "Are I all Better Drivers than Average? Self Perception and Biased Behaviour," CEPR Working Paper 3603, Oct. 2002.

Chevalier, J. and C. Avery, "Herding over the Career," *Economics Letters* 63, 1999, 327-333.

Cho, I.K. and D. Kreps, "Signalling Games and Stable Equilibria," *Quarterly Journal of Economics* 102, 1987,179-221.

Cho, I.K. and Sobel, J., "Strategic Stability and Uniqueness in Signalling Games," *Journal of Economc Theory* 50, 1990, 381-413.

Daniel, Kent, David Hirshleifer and Avanidhar Subrahmanyam, "Investor Psychology and Security Market Under- and Overreactions," *Journal of Finance* 53(6), Dec. 1998, 1839-1885.

Dewatripont, Mathias, Ian Jewitt and Jean Tirole," The Economics of Career Concerns, Part I: Comparing Information Structures," *Review of Economic Studies* 66, 1999, 183-198.

Fischhoff, Baruch., Paul Slovic and Sarah Lichtenstein, "Knowing with Certainty: The Appropriateness of Extreme Confidence," *Journal of Experimental Psychology: Human Perception and Performance* 3, 1977, 552-564.

Fudenberg, Drew and Jean Tirole, *Game Theory*, London, MIT Press, 1992.

Gilat, Levy, "Anti-herding and Strategic Consultation," *European Economic Review* 48, 2004, 503-525.

Gilpin, Kenneth N., "Market Insight; The Numbers May Change, But Not the Optimism." *New York Times* 28 Mar. 1999.

Glaser, Markus, Markus Nöth and Martin Weber, "Behavioural Finance," in Derek J. Koehler and Nigel Harvey, eds., *Blackwell Handbook of Judgment and Decision Making*, Oxford, Blackwell Publishing, 2004.

Gregorovich, Andrew, "Chernobyl Nuclear Catastrophe: Ten Years After April 26, 1986," *Forum: A Ukrainian Review* 94, Spring 1996.

Hirschleifer, David, "Investor Psychology and Asset Pricing," *Journal of Finance* 61(4), Aug 2001, 1533-1597.

Johnson, Dominic D. P., *Overconfidence in War: The Havoc and Glory of Positive Illusions*, Cambridge, Harvard University Press, 2004.

Koriat, Asher, Sarah Lichtenstein and Baruch Fischhoff, "Reasons for Confidence," *Journal of Experimental Psychology: Human Learning and Memory* 6, 1980, 107-118.

Kreps, David and Robert Wilson, "Reputation and Imperfect Information," *Journal of Economic Theory* 27, 253-279.

Kruger, Justin and David Dunning, "Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self Assessments," *Journal of Personality and Social Psychology* 77, 1999, 1121-1134.

Mailath, George J., "Incentive Compatibility in Signaling Games with a Continuum of Types," *Econometrica* 55, 1349-1365.

Nickerson, Raymond S., "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises," *Review of General Psychology* 2, 1998, 175-220.

Oaksford, Mike and Nick Chater, "A Rational Analysis of the Selection Task as Optimal Data Selection," *Psychological Review* 101, 1994, 608-31.

Plous, Scott, *The Psychology of Judgment and Decision Making*, London, McGraw-Hill, Inc., 1993.

Rajan, Raghuram, "A View of the Liquidity Crisis," speech, Chicago, 2008.

Ramey, Garey, "D1 Signaling Equilibria with Multiple Signals and a Continuum of Types," *Journal of Economic Theory* 69, 796-821.

Riley, John, "Informational Equilibria," *Econometrica* 47, 331-359.

Samuelson, Robert J., "Who's to Blame? Why Capitalists are Capitalism's most Dangerous Enemies," *Newsweek*, URL: http://www.newsweek.com/id/98099/output/print, updated 23 Jan. 2008.

Svenson, Ola, "Are I all less risky and more skillful than our fellow drivers?" *Acta Psychologica* 47, Feb. 1981, 143–148.

Van den Steen, Eric, "Rational Overoptimism and (Other Biases)," *American Economic Review* Sept. 2004, 1141-1151.

Wason, P. C., "On the Failure to Eliminate Hypotheses in a Conceptual Task," *Quarterly Journal of Experimental Psychology* 12, 1960, 129-40.

Weston, Drew, Pavel S. Blagov, Keith Harenski, Clint Kilts and Stephan Hamann, "Neural Bases of Motivated Reasoning: An fMRI Study of Emotional Constraints on Partisan Political Judgement in the 2004 U.S. Presidential Election," *Journal of Cognitive Neuroscience* 18:11, 2006, 1947-1958.