Department of Economics

Psychology and Economics

Jean Tirole

Self-Confidence:
Intrapersonal Strategies

Roland Bénabou
and
Jean Tirole

EIB LECTURE SERIES

EUROPEAN UNIVERSITY INSTITUTE

Printed in Italy in October 2000
European University Institute
Badia Fiesolana
I – 50016 San Domenico (FI)
Italy

SELF-CONFIDENCE:

INTRAPERSONAL STRATEGIES[1]

Roland Benabou[2] Jean Tirole[3]

First version: June 1999
This version: December 1999

[1]We are grateful for helpful comments and discussions to Dilip Abreu, Olivier Blanchard, Isabelle Brocas, Dan Gilbert, David Laibson, George Loewenstein, Andrew Postlewaite, Marek Pycia, Matt Rabin, Julio Rotemberg and participants at the GREMAQ-CEPR conference on Economics and Psychology (Toulouse, June 1999) and the NBER Conference on Macroeconomics and Individual Decision Making (Cambridge, November 1999).

[2]Princeton University, NBER, CEPR and IRP.
[3]IDEI and GREMAQ (UMR 5604 CNRS), Toulouse, CERAS (URA 2036 CNRS), Paris, CEPR, and MIT.

Abstract

This paper analyzes the self-identification process and its role in motivation. We build a model of self-confidence where people have imperfect knowledge about their ability, which in most tasks is a complement to effort in determining performance. Higher self-confidence thus enhances motivation, and this creates incentives for the manipulation of self-perception. An individual suffering from time-inconsistency may thus want to enhance the self-confidence of his future selves, so as to limit their procrastination. The benefits of confidence-maintenance must, however, be traded off against the risks of overconfidence (inappropriate tasks being pursued). Moreover, rational inference implies that the individual cannot systematically fool himself. A first application of the model is self-handicapping: to avoid a negative inference about their ability, people may deliberately impair their performance, or choose overambitious tasks. Another application is selective memory or awareness management: people are (endogenously) more likely to remember or consciously acknowledge their successes than their failures. This, in turn, helps explain the widely documented prevalence of self–serving beliefs –that is, the fact that most people have overoptimistic assessments of their own abilities and other desirable traits. We analyze the workings of this "psychological immune system" and show that it typically leads to multiple equilibria in cognitive strategies, self confidence, and behavior. Moreover, while active self-esteem maintenance can improve ex-ante welfare, it can also be self-defeating. Systematically "looking on the bright side", avoiding "negative" thoughts and people, etc., can thus be beneficial in certain environments; but in other circumstances one can only lose by playing such games with oneself, and it would be better to always "accept who you are" and "be honest with yourself ".

Keywords: self–confidence, self–esteem, motivation, time-inconsistency, self-control, self–deception, memory, psychology and economics.

JEL Classification: A12, C70, D60, D91, E21, J22, J24.

"Believe what is in the line of your needs, for only by such belief is the need fulfilled... Have faith that you can successfully make it, and your feet are nerved to its accomplishment".

William James, *Principles of Psychology.*

"I have done this, says my memory. I cannot have done that, says my pride, remaining inexorable. Finally–memory yields."

Friedrich Nietzsche, *Jenseits von Gut und Böse.*

"I had during many years followed the Golden Rule, namely, that whenever a published fact, a new observation or thought came across me, which was opposed to my general results, to make a memorandum of it without fail and at once; for I had found by experience that such (contrary and thus unwelcome) facts and thoughts were far more apt to escape from memory than favorable ones."

Charles Darwin. *The Life of Charles Darwin,* by Francis Darwin.

"Repression is automatic forgetting."

H. P. Laughlin, *The Ego and Its Defenses.*

1 Introduction

Throughout the history of psychology, the maintenance and enhancement of self-esteem has been identified as a fundamental human impulse. Understanding the process of self-identification and the nature of self-knowledge is crucial to studies of motivation, affect, and social interactions.

The purpose of this paper is to bring these concerns into the realm of economic analysis, and show that this has important implications for how agents process information and make decisions in all spheres of life. Conversely, the tools of economic modelling can help shed light on of a number of apparently irrational behaviors documented by psychologists.

The paper thus proposes a formal framework, the *self-confidence model* (SCM), which unifies different themes emphasized in the social psychology literature and draws out their economic implications. The self-confidence model rests on the following hypotheses:

Hypothesis 1 (imperfect information). *People have imperfect knowledge about their enduring personal characteristics, or more generally about the eventual costs and payoffs of their efforts.*

This assumption is rather uncontroversial. The psychology literature generally views introspection as quite inaccurate (Nisbett and Wilson 1977) and stresses that learning about oneself is an ongoing process of self-identification. While people aspire to succeed in their private and work lives, gain social acceptance, quit smoking, or play the piano better, they are uncertain as to their ability to reach their goals. They learn about their self and get to know their abilities over time. Yet, this process may never converge, as the self is constantly changing.[1] For one thing, personal characteristics may change over time; perhaps more importantly, goals change over time: rewards attached to success in one favorite dimension may decrease, the private cost of pursuing past goals may increase, and learning through interactions with others may suggest new objectives. Changing goals means moving to activities over which the self-knowledge is less accurate.

Last but not least, the environment which determines the return to effort is typically changing. While the term "ability" usually refers to an individual-specific attribute, for our purposes it can also be specific to the task (long-run expected payoff, how difficult or enjoyable it will be to complete), or to the match between the two (consumer tastes, presence or lack of discrimination against women or minorities). In interpreting the model we shall generally focus on the first case, but it will be clear from the analysis that all three are formally equivalent.

Hypothesis 2 (cognitive ability). *The individual is a rational (Bayesian) information processor who does not fool himself on average.*

While familiar in psychology (since e.g. Festinger 1954), this assumption is not as widely accepted as in economics. We should therefore discuss it in some detail.

In the ongoing process of self-identification, the individual uses three sources of information: the conscious memory of past acts and performances (successful or failed goal completion) by himself and comparable others, the retrieval of unconscious events, and the feedback received through others' assessments or their attitudes towards him. There is a good amount of evidence that, while probably not an accurate statistician, the individual updates his beliefs according to broad Bayesian principles. At the same time, it is also widely recognized that information acquisition and belief updating are subject to self-serving biases. We first discuss the rational side of inference, then turn to the evidence suggesting a less rational side, or at least a *motivated* one.

– *Attribution theory* (Heider 1958) emphasizes the distinction between temporary (situational) and enduring (dispositional) characteristics.[2] In economics parlance, the individual filters out noise in order to extract information about his type from past performance. In so doing, he adopts

[1] E.g., Rhodenwalt (1986).

[2] E.g. Gilbert and Silvera (1996).

an "applied scientist's view": an effect is attributed to causes with which it covaries (e.g., Kelley 1972). Similarly, variations in self-esteem are much more likely when the individual changes social environment, as would be expected from the fact that he then becomes exposed to new sources of self-relevant information.

– In the *social comparison process* (Festinger 1954), individuals rationally assess their own ability by comparing their performance with that of people facing similar conditions (familial, cultural, educational, situational,...). In the language of economics, they use "relative performance evaluation", or "benchmarking", for self-evaluation. A good performance by others in one's reference group is thus generally detrimental to self-esteem, as it leads to a feeling of inferiority.[3] Conversely, the benchmarking process implies that some comfort is derived when others experience adversity.[4] Indeed, one key purpose of group therapy is to provide people with low self-esteem, a depressive mood, or substance-abuse problems, with information that they are not exceptional subjects. For example, Alcoholic Anonymous aims at demonstrating to its members, through group interaction, that other alcoholics are not hopeless bums but rather good and attractive people who have experienced hardship and try to escape their condition.

– Relatively sophisticated updating also applies to the interpretation of others' behaviors and assessments of oneself, such as praise and criticism. For example, it a standard observation that, when extracting self–relevant information, a person takes into account not only what others say or do, but also their intentions. Also in the area of interpersonal relations, it is known that the individual discounts his performance when being helped by others. This is why overhelping is an effective way of spoiling a person's perceived ability, especially when help does less to performance than observers think it does (Gilbert and Silvera 1996).

While the earlier literature emphasized rationality and information-seeking in the process of self-identification, later work included many attempts to document the less rational side of human inference, and in particular the prevalence of *self-serving biases*. For instance, a substantial body of evidence suggests that people make internal attributions for successes and external ones for failures (e.g., Zuckerman 1979).[5] More generally, they tend to overestimate their abilities and other desirable traits, both in absolute terms and relative to others (e.g., Taylor and Brown 1988).

[3] The exception is when the high performers are related to the individual, allowing him to "bask in their reflected glory". Thus, people are pleased when their cousin is a great pianist, or their neighbor a movie star; presumably, the positive inferences which can be drawn from their being associated with successful others then outweigh the bad news of not having matched their performance.

[4] This was noted by classical authors such as La Rochefoucauld, Kant and Goethe (the following cites are drawn from Heider 1958): *"The presence of the wretched is a burden to the happy; and, alas, the happy still more so to the wretched"* (Goethe, Stella, Act 3); it can help explain the existence of Schadenfreude: *"In the adversity of our best friends we always find something that is not displeasing"* (La Rochefoucauld, 1665, Reflexions, Paris-Garnier).

[5] Why they would want to do so in a social context is obvious. The interesting question is why they may bias their own inference process.

They also rate their own probabilities to be above average for favorable future events, and below average for unfavorable ones; the more controllable these events through their future actions, the more so (Weinstein 1980). Similarly, most nondepressed individuals seem to overestimate the extent to which they have control over outcomes, while depressed subjects have much more realistic assessments (Alloy and Abrahamson 1979).

Baumeister (1998) provides a very good review of these many facets of the general phenomenon of *self-deception*, which has puzzled philosophers throughout the ages. Indeed, while it may serve well to enhance one's self-esteem, this process seems at odds with standard principles of statistical updating, or even reasonable approximations thereof. It has thus been argued that self-deception fails to explain how an individual may end up believing both an event and its contrary (Sartre 1953), and that it is a "concept in search of a phenomenon" (Gur and Sackeim 1979).[6]

Because of our insistence on hypothesis 2, we will need to address self-deception. Section 3 will attempt to reconcile the esteem-maintenance ("hot") and cognition ("cold") features of attribution. The Bayesian approach will be seen to be quite compatible with the prevalence of motivated, self-serving biases, and to deliver a much richer set of results than a purely naive one where agents are viewed as completely unaware of the processes by which they arrive at their current beliefs. In particular, there may multiple, intrapersonal equilibria, and self-deception may ultimately be hurtful. Later on in Section 4 we shall relax Hypothesis 2 by allowing individuals to be imperfect Bayesians, and show that the distinguishing features of the rational approach persist as long as beliefs are not too naive.

Hypothesis 3 (relevance of self-knowledge). *Ability and effort are complementary factors in the production of performance.*

We posit that self-confidence affects the incentive to undertake actions. Namely, the higher one's perceived ability in an activity, the stronger the incentive to undertake this activity and persevere until its completion. Technically, ability and effort (which, together with situational factors, i.e., luck, determine performance) are complementary factors of production.

As demonstrated by the earlier quote from James (1890), this complementarity has long been familiar in psychology.[7] Gilbert and Cooper (1985) note that *"the classic attributional model of the causes of behavior argues that the individual's effort combines multiplicatively with his abilities to determine outcomes. Thus the well-known conceptual equation: $(E \times A) \pm TD = B$, in which effort times ability, plus or minus task difficulty equals the behavioral outcome."*

Complementarity is also implicit in Bandura's (1977) self-efficacy theory, according to which *"beliefs of personal efficacy constitute the key factor of human agency."*[8] Were the production function additively separable, beliefs about one's capabilities would be irrelevant to the choice of effort. Complementarity between perceived ability and effort is also consistent with the standard observations that when individuals expect to fail, they fail quite effectively, and that failure leads to failure more readily for individuals characterized with low self-esteem (Salancik 1977).

Remark 1: There are also instances, however, where perceived ability and effort are *substitutes;* we shall consider this variant of the model in Section 5.1. This typically occurs when the payoff is of a zero–one, or "pass–fail" nature, such as in obtaining a diploma, making a sale, being hired or fired (tenure, partnership), as well as in marriage and divorce. Since a high perceived ability may then increase the temptation to exert low effort, this version of the model allows us to account for what psychologists refer to as *"defensive pessimism":* the fact that people sometimes tend to minimize, rather than aggrandize, their previous accomplishments and expectations of future success (conditional on a given level of effort).

Remark 2: In our model, self-confidence matters through its impact on behavior. This "executive function" view of the self, while pervasive in psychology, is by no means the only root of selfhood (Baumeister 1998). A common and complementary root views the self as backward- rather than forward-looking, with self-esteem generating purely affective, rather than functional, benefits. Moping over one's failures and inadequacies or glorifying in one's triumphs when lying in bed late at night is an example.[9] Sometimes referred to by psychologists as "self–affirmation," this alternative view of why self-esteem matters may be best described in economists' terms by labeling self-esteem as a *consumption good,* as opposed to as a source of motivation in the executive function view. The general approach of putting beliefs directly into the utility function was first used in Akerlof and Dickens' (1982) well–known paper on cognitive dissonance, and more recently (in different contexts) by Rabin (1995), Caplin and Leahy (1999), Weinberg (1999) and Köszegi (1999)). We shall discuss it in Section 5.2, and show that several of the main insights from the SCM carry over to this alternative setting.

The combination of Hypotheses 1 and 3 has an important behavioral consequence which, as we shall see, is relevant to a wide variety of situations.

Corollary (manipulative behavior). *The individual's imperfect self-knowledge creates, for himself and/or others, incentives to manipulate information concerning the self.*

[6] For recent discussions of self-deception, see Gilbert and Cooper (1985), Mele (1999) and Elster (1999). Gilbert and Cooper note for instance that we are all insightful "naive psychologists", well aware of human tendencies to be self-serving, and argue that self-deceptive strategies must produce self-inflated assessments "in ways that seem reasonable, accurate and fair".

[7] See also Heider 1958, or Darley and Goethals 1980. In economics, Dewatripont et al. (1999a,b) make substantial use of this complementarity in their analysis of career concerns incentives.

[8] See also Deci (1975).

[9] See Baumeister (1998) for other examples.

In standard decision theory, a rational individual has no incentive to manipulate his beliefs: accurate information is always better than a coarser one. By contrast, in a game-theoretic context it is well-known that players' divergent interests generally give them incentives to manipulate each other's information.[10] Hypothesis 3 offers a clue as to why manipulation can be effective in the SCM: a higher self-confidence enhances the individual's motivation. Hence, anyone with a vested interest in his performance has an incentive to build up and maintain the individual's self-esteem. This can arise in two types of situations.

First, the manipulator could be another person (parent, teacher, spouse, friend, colleague, boss or relation) who is eager to see the individual "get his act together", or otherwise succeed in the task at hand. The spillover effect of self- esteem then provides the motivation for manipulative behavior. Such *interpersonal strategies* are studied in Benabou and Tirole (1999).[11]

Second, for an individual suffering from time inconsistency, the current self has a vested interest in maintaining and enhancing the self-confidence of future selves, so as to counter their natural tendency to procrastinate. As in Carrillo and Mariotti (1997), the self therefore has an incentive to manipulate the beliefs of future selves. Such *intra-personal strategies* are the topic of this paper.

Inter- and intrapersonal strategies will be analyzed through the lens of standard game theory. This modeling choice is motivated by simplicity; for our purpose, strategies could alternatively be the outcome of learning, evolution, education or instinct.

2 The self-confidence model: basics and a first application

2.1 The procrastination of effort

"Had I been less definitively determined to start working, I might have made an effort to begin right away. But because my resolve was absolute and, within twenty-four hours, in the empty frames of the next day where everything fit so well since I was not yet there, my good resolutions would easily be accomplished, it was better not to choose an evening where I was ill disposed for a beginning to which, alas! the following days would turn out to be no more propitious."

Marcel Proust, *Remembrance of Things Past.*

1. Preferences and decisions. Consider a risk-neutral individual with a relevant horizon equal to three periods: $t = 0, 1, 2$. At date 0, he selects an action; for full generality, we do not specify the date 0 action space yet. It suffices at this stage to know that the date 0 action determines

[10] See, e.g., Fudenberg and Tirole (1991) and Osborne and Rubinstein (1994) for overviews of noncooperative game theory with imperfect information.

[11] The spillover may exist even under pure altruism, as long as the manipulated individual suffers from time inconsistency.

(possibly among other things) the individual's *information structure* at date 1. The simplest date 0 action is thus the choice of a date 1 information structure, that is, the amount and type of date 1 self-relevant information acquired at date 0. Alternatively, the information acquired at date 0 may be derived from the outcome of some activity pursued at that date.

At date 1, the individual may undertake the date 1 activity (exert effort) or not (exert no effort). The first option involves an effort cost c at date 1, the second one involves no cost. If successful, the activity generates benefit V to the individual at date 2; failure generates no benefit.

The probability of success as perceived by the individual at date 1 is denoted $\theta \in [0, 1]$. That is, θ stands for the individual's date 1 *self-confidence* (or self-esteem).

Last, we assume that the individual exhibits hyperbolic discounting and therefore suffers from the standard *time-inconsistency* problem. There is indeed considerable experimental evidence that individual choices exhibit a "salience of the present", in the sense that discount rates are much lower at short horizons than at more distant ones.[12] Formally, the intertemporal payoff perceived by the individual's self at date 1 is

$$u_1 + \beta\delta u_2 = -c + \beta\delta\theta V \tag{1}$$

when undertaking the activity (and 0 when not undertaking it). By contrast, the date 0 self's intertemporal payoff is

$$u_0 + \beta\left[\delta u_1 + \delta^2 u_2\right] = u_0 + \beta\delta\left[-c + \delta\theta V\right] \tag{2}$$

if the date 1 self undertakes the activity (and u_0 when he does not), where u_0 is the date 0 instantaneous payoff. We assume that $\beta < 1$, which implies that the date 0 self ("Self 0") is concerned about the date 1 self's "excessive" preference for the present. Indeed, the date 1 self ("Self 1") undertakes the activity if and only if

$$\beta\delta\theta V \geq c.$$

that is, if his self-confidence exceeds $c/\beta\delta V$. By contrast, from the point of view of Self 0, the activity should be undertaken whenever

$$\delta\theta V \geq c,$$

that is, when Self 1's self-confidence exceeds $c/\delta V$. This is the standard result that the date 0 self

[12] See Ainslie (1992) and references therein for the evidence; Strotz (1956), Phelps and Pollack (1968), Loewenstein and Prelec (1992), Laibson (1994, 1997) and O'Donoghue and Rabin (1999) for formal models and economic implications; and Mulligan (1996) for a dissenting view.

is concerned about date 1 procrastination. It is worth noting at this point that while we focus the exposition on the case where θ (the individual's intrinsic ability) is unknown, it could equally well be the "survival" probability δ, or the task's difficulty, measured by the cost of effort c. All that matters for our theory is that the individual be uncertain of the long term *return to effort* $\theta\delta V/c$ which he faces.

2. Payoffs. Two date 0 actions in general differ in their date 0 payoff (u_0) as well as in their impact on date 1 self-knowledge. This section focuses on the latter, that is on the date 0 preferences over information structures about the self at date 1.

Without loss of generality, we can define θ as the expectation of the individual's ability at date 1 (given risk-neutrality, only the expectation matters). An action that generates cumulative distribution function $F(\theta)$ over self-1's perceived expected ability $\theta \in [0,1]$ generates a date 1 flow payoff, *as viewed from date 0*, equal to $\beta\delta$ times

$$\int_{c/\beta\delta V}^{1} (\delta\theta V - c)\, dF(\theta). \tag{3}$$

Ignoring their impact on u_0, two date 0 actions can therefore be compared from the point of view of their impact on date 1 information, namely on the posterior distribution $F(\theta)$.

2.2 Confidence maintenance vs. overconfidence

It is well-known from the work of Carrillo and Mariotti (1997) that, in the presence of time inconsistency (TI), Blackwell garblings of information may increase the current self's payoff. This result can be usefully reinterpreted, and further developed, in our context.

Consider a date 0 self who chooses between two information structures for the date 1 self. In the finer information structure, Self 1 learns his ability, drawn from cumulative distribution $F(\theta)$. In the coarser information structure, Self 1 learns nothing that Self 0 does not know, and therefore has expected ability equal to the mean of the distribution F : $\bar{\theta}_F \equiv \int_0^1 \theta dF(\theta)$.

Let us first assume that when acquiring no information, Self 1 undertakes the activity: $\bar{\theta}_F > c/\beta\delta V$. The value of information for Self 1 as assessed by Self 0 is therefore $\beta\delta$ times

$$\mathcal{I}_F \equiv \int_{c/\beta\delta V}^{1} (\delta\theta V - c)\, dF(\theta) - \left(\delta\bar{\theta}_F V - c\right) = \mathcal{G}_F - \mathcal{L}_F, \tag{4}$$

where

$$\mathcal{G}_F \equiv \int_{0}^{c/\delta V} (c - \delta\theta V)\, dF(\theta), \tag{5}$$

$$\mathcal{L}_F \equiv \int_{c/\delta V}^{c/\beta\delta V} (\delta\theta V - c)\, dF(\theta). \tag{6}$$

$\mathcal{G}_F$ stands for the gain from being informed, which arises from the fact that better information reduces the risk of *overconfidence* on the part of Self 1. Overconfidence occurs when the individual's ability is lower than $c/\delta V$ but he is unaware of it, and undertakes the activity. $\mathcal{L}_F$ stands for the loss from being informed, which may depress the individual's self–confidence: if he learns that θ is in some intermediate range, $\theta \in [c/\delta V, c/\beta\delta V]$, he will procrastinate at date 1 even though, ex–ante, is was optimal to work. Information is thus detrimental to the extent that it creates a risk that the individual will fall into this time-inconsistency (TI) region. If this effect is strong enough, $\mathcal{L}_F > \mathcal{G}_F$, the individual will prefer to remain uninformed. More generally, note that $\mathcal{I}_F$ is lower, the lower is β.

An example where $\mathcal{I}_F < 0$ is when initial self-confidence is high enough that the support of the distribution F lies in $[c/\delta V, 1]$. If he knew his ability , Self 0 would then always want Self 1 to exert effort. In this case, $\mathcal{G}_F = 0$: there is no risk that Self 1 will be overconfident, hence no gain to information. There is, however, a loss if the distribution F puts some weight in the TI region. More generally, we can define *confidence maintenance*, or *self-esteem maintenance*, as the individual's reluctance to accept free information, or equivalently as his willingness to accept Blackwell garblings. By contrast, in the absence of time inconsistency ($\beta = 1$) we have $\mathcal{L}_F = 0$; as usual in classical decision theory, information is always valuable, $\mathcal{I}_F \geq 0$.

To sum up, the concern about overconfidence relates to the possibility that poor self-knowledge leads one to undertake (or pursue) an activity when one's actual ability is low ($\theta < c/\delta V$). By contrast, confidence maintenance is motivated by the concern that better information may put the next self in the TI region of intermediate self-confidence ($c/\delta V < \theta < c/\beta\delta V$). The overconfidence effect calls for more information, confidence maintenance for less. This tradeoff has been noted by empirical researchers. For instance, Leary and Downs (1995) summarize the literature by noting that:

a) "persons with high self esteem perform better after an initial failure and are more likely to persevere in the face of obstacles";

b) "high self-esteem is not always functional in promoting task achievement. People with high self-esteem may demonstrate non-productive persistence on insoluble tasks, thereby undermining their effectiveness. They may also take excessive and unrealistic risks when their self-esteem is threatened".

To understand the last statement, let us now turn to the case where $\bar{\theta}_F < c/\beta\delta V$. It is clear that information is now always valuable:

$$\mathcal{I}_F = \int_{c/\beta\delta V}^{1} (\delta\theta V - c)\, dF(\theta) > 0. \tag{7}$$

This is simply due to the fact that Self 1 always exerts (weakly) less effort than Self 0 would like

him to. Information can thus only help the individual, by potentially restoring his self-confidence and motivation. Accordingly, $\mathcal{I}_F$ is now *higher*, the greater the time-consistency problem. In such situations the individual will avidly seek information about his ability, and his choices of tasks and social interactions will have the nature of "*gambles for resurrection*" of his self-esteem.

Putting together the different cases, we see that the *value of information* (the amount, positive or negative, that the person is willing to pay in order to know θ) *is not monotonic with respect to initial self confidence* (more formally, with respect to a higher distribution of θ, in the sense of first-order stochastic dominance). Indeed, for someone with confidence so low that $\bar{\theta}_F < c/\beta\delta V$, $\mathcal{I}_F$ is always positive and increasing with respect to (stochastic) increases in θ.[13] For an individual with confidence such that $F(c/\delta V) = 0$ but $F(c/\beta\delta V) < 1$, we saw that $\mathcal{I}_F$ is always negative. Finally, for a person so self-assured that $F(c/\beta\delta V) = 0$, procrastination is not a concern (as if β were equal to 1), but neither is overconfidence: $\mathcal{I}_F = \mathcal{G}_F = \mathcal{L}_F = 0$. Therefore, there must exist some intermediate range where $\mathcal{I}_F$ first declines and becomes negative, then starts increasing again towards zero.

2.3 What types of individual are most eager to maintain their self-confidence?

To keep forging intuition about these two basic effects, let us consider an individual who faces a date 0 prior distribution $F(\theta)$ over his ability, and analyze his eagerness to learn his type. As we noted, his willingness to learn is always positive if ignorance leads to inactivity at date 1. Suppose instead that ignorance leads Self 1 to undertake the activity ($\bar{\theta}_F > c/\beta\delta V$). From (4), Self 0's willingness to pay for information is ($\beta\delta$ times)

$$\mathcal{I}_F = \int_0^{c/\beta\delta V} (c - \delta\theta V)\, dF(\theta) = \delta V \int_0^{c/\beta\delta V} F(\theta)\, d\theta - \left(\frac{1-\beta}{\beta}\right) cF\left(\frac{c}{\beta\delta V}\right). \tag{8}$$

Abstracting for the moment from any cost attached to learning or not learning the true ability, Self 0 is willing to accept the information ($\mathcal{I}_F \geq 0$) if and only if

$$\int_0^{c/\beta\delta V} \frac{F(\theta)}{F(c/\beta\delta V)} d\theta \geq \left(\frac{1-\beta}{\beta}\right)\left(\frac{c}{\delta V}\right). \tag{9}$$

Let us now compare two individuals with different degrees of initial self-confidence and ask: which of the two is more concerned about confidence maintenance, that is, less willing to receive or listen to information? We denote the distributions over abilities as $F(\theta)$ and $G(\theta)$, their densities as $f(\theta)$ and $g(\theta)$, and their means as $\bar{\theta}_F$ and $\bar{\theta}_G$. To make confidence maintenance meaningful, we

[13] Rewrite (7) as $\mathcal{I}_F = \int_0^1 1_{\{\theta \geq c/\beta\delta V\}} (\delta\theta V - c)\, dF(\theta)$, where $1_{\{\cdot\}}$ denotes the indicator function, and note that the integrand is increasing in θ.

assume that $\bar{\theta}_F > \bar{\theta}_G > c/\beta\delta V$. Let us now define greater or lesser self-confidence:[14]

Definition 1 (Comparison of self-confidence levels). *An individual with distribution F over ability θ has higher self-confidence than an otherwise identical individual with distribution G if the likelihood ratio is monotonic, meaning that:*

$$\frac{f(\theta)}{g(\theta)} \text{ is increasing in } \theta.$$

The monotone likelihood ratio property (MLRP) implies that $F(\theta) < G(\theta)$ for all $\theta \in (0,1)$ (first-order stochastic dominance), that $\bar{\theta}_F > \bar{\theta}_G$, and that $F(\theta)/G(\theta)$ is increasing on $(0,1)$.

From condition (9), the individual with the higher self-confidence is more concerned about confidence maintenance (in the sense of being willing to accept information about his ability for a smaller set of parameters) if and only if

$$\int_0^{c/\beta\delta V} \frac{F(\theta)}{F(c/\beta\delta V)} d\theta \leq \int_0^{c/\beta\delta V} \frac{G(\theta)}{G(c/\beta\delta V)} d\theta. \tag{10}$$

It is easy to show that the monotone likelihood ratio property implies (10).[15] Intuitively, the individual with the greater initial self-confidence has more to lose from information, and is therefore the most *insecure*. Finally, recalling that the information always has positive value when no effort is undertaken in its absence, we have shown:

Proposition 1 *If an individual prefers not to receive information in order to preserve his self-confidence, so will any individual with higher initial self-confidence: if $\mathcal{I}_G < 0$ for some distribution G over θ, then $\mathcal{I}_F < 0$ for any distribution F such that the likelihood ratio f/g is increasing.*

2.4 Self-handicapping

Starting with the early work of Berglas and Jones (1978), a well-documented and puzzling phenomenon is that people sometimes create obstacles to their own performance.[16] In principle, they should want to perform up to the best of their ability. Yet, in experiments, subjects with fragile self-confidence may accept taking performance-impairing drugs before an intelligence test. In real life, people withhold effort, prepare themselves inadequately, or drink alcohol before undertaking

[14] We are grateful to Ian Jewitt for drawing our attention to the relation between this problem and the total positivity literature.

[15] The MLRP implies that $F(\theta)/G(\theta) \leq F(c/\beta\delta V)/G(c/\beta\delta V)$, for all θ in $[0, c/\beta\delta V]$, or equivalently $F(\theta)/F(c/\beta\delta V) \leq G(\theta)/G(c/\beta\delta V)$.

[16] See, e.g., Arkin and Baumgardner (1985), Baumeister (1998), Fingarette (1985), Gilovich (1991), and Hull and Young (1983).

a task. Test anxiety and "choking" under pressure during an athletic or artistic performance are yet other common forms of self-handicapping.

It has long been suggested that self-handicapping is a self-esteem maintenance strategy. According to this view, people are willing to sacrifice current performance for a protection against a loss in self-esteem. In this respect, the self-setting of *overambitious goals* or choice of *overambitious tasks* may also be viewed as a form of self-handicapping: people impose a current cost on themselves (the loss in expected current payoff when selecting an inappropriate task rather than a more adequate one, the likely cost of losing face in front of others, etc.), but they reduce the probability of a large negative inference about their self (in the limit, an infeasible task generates no inference in case of failure).

To be certain, self-handicapping behavior is more complex than just self-esteem maintenance, because it involves both intrapersonal and an intrapersonal (self-presentation) strategies. As Baumeister (1998) notes, self-handicapping *"has clear cognitive implications, especially regarding attributions: by self-handicapping, one can forestall the drawing of unflattering attributions about oneself. Self-handicapping makes failure meaningless, and so if people think you are intelligent the upcoming test cannot change this impression."* In particular, people apparently self-handicap more in public situations (Kolditz and Arkin 1982). They may then reap a double dividend, since self-handicapping provides an excuse for poor performance both to themselves and to others.[17]

This section will not address the debate concerning the relative influence of the private and public selves in generating self-handicapping. Rather, it verifies the logical consistency of the self-esteem rationale for self-handicapping behaviors. The combination of the hypotheses of complementarity between effort and ability and of time inconsistency embodied in the SCM provides the mechanism through which an individual may want to self-handicap.

To analyze this issue more formally, let us compare two date 0 actions. Assume that the efficient action from the point of view of date 0 (that is, the action that maximizes the date 0 flow payoff) is also more informative. That is, it reveals the true ability θ, while the inefficient action is uninformative, and thus leaves the individual at date 1 with beliefs still described by the prior distribution $F(\theta)$. As earlier, let us assume that the expectation of θ given distribution $F(\theta)$ exceeds $c/\beta\delta V$. Last, let $h_0(F) \geq 0$ denote the date 0 cost of choosing the inefficient, "self-handicapping" action.

Generalizing equation (8), the individual will self-handicap if and only if $-\beta\delta\mathcal{I}_F \geq h_0(F)$, or:

$$\left(\frac{1-\beta}{\beta}\right)c - \delta V\left[\int_0^{c/\beta\delta V} \frac{F(\theta)}{F(c/\beta\delta V)}d\theta\right] \geq \frac{h_0(F)}{\beta\delta F(c/\beta\delta V)}. \tag{11}$$

[17] Furthermore, under imperfect memory, providing excuses for one's poor performance to others conceivably may provide excuses to oneself and maintain self-esteem, to the extent that others may behave differently in the future.

We saw earlier that $I_F < 0$ when $\beta < 1$ and the individual is not too concerned about the risk of overconfidence (exerting effort when it is suboptimal). In such situations he will choose to self-handicap, provided the costs of doing so are not too large.

Let us now try to go further and compare two individuals with different degrees of self-confidence, as defined by distributions F and G with the MLRP (as in Definition 1). It may seem a priori reasonable that the cost of self-handicapping should be higher for the individual with the higher self-esteem:

$$h_0(F) \geq h_0(G). \tag{12}$$

If we think of self-handicapping as an action that inhibits performance in a task, the inhibition cost would seem to be higher, the higher the subject's perceived ability. For instance, the higher the probability of success in a task, the higher the cost of botching the task. There are also plausible examples, however, where (12) is reversed. Such is the case for instance if the task is of a pass/fail nature (for example an exam), but may also yield ego-relevant information (e.g., the grade). A more self-assured individual can than more safely engage in self-handicapping, so to as the make the grade uninformative, without much risk of actually failing.

Turning now to the analysis of (11), we saw earlier that the MLRP implies that the expression on the left-hand side is larger than if F were replaced with G : those who are initially more self-confident have *more to lose*. However, the MLRP also implies that $F(c/\beta\delta V) \leq G(c/\beta\delta V)$, which tends to make the right-hand side of (11) also larger under F than under G : the more self-confident are *less likely* to receive bad news, and this reduces the return on investing h_0 in "non-information". This effect is compounded when (12) holds, and weakened in the reverse case.

Thus, in general, one cannot conclude whether it is the individual with higher of lower self-confidence which is more likely to self-handicap. This ambiguity is fundamentally linked to the *non-monotonicity of value of information* $\mathcal{I}_F$, which was noted as at the end of Section 2.2. While the MLRP ensures that the *sign of* $\mathcal{I}_F$ varies monotonically with initial self-confidence, what matters for self-handicapping is the *absolute amount* which the individual is willing to pay to suppress the information.

It is interesting to note that while the experimental psychology literature has also not reached a firm conclusion on whether high or low self-confidence people are the most likely to be defensive of their egos, the preponderance of evidence seems to be in favor of the former hypothesis (which would at least be consistent with Proposition 1). Thus Greenier et al. (1995) contrast "humanistically oriented theories (e.g., Rogers 1961), according to which high self-esteem individuals's feelings of self-worth are built on solid foundations that do not require continual validation", with experimental research showing that "high self-esteem individuals are the more likely to dis-

play self-serving attributions, self-handicap to enhance the potentially positive implications of good performance, set inappropriately risky goals when ego-threatened, and actively create less fortunate others with whom they can compare favorably."

Remark: What about the observation that individuals with very low self-esteem (with expected ability lower than $c/\beta\delta V$ in our model) may also engage in behaviors such as drinking? In our purely intrapersonal-strategy model they have no such incentive. There are two possible interpretations for such a behavior. The first is that this is a type of self-handicapping motivated by self-presentation: being drunk provides an "excuse" for failure or for the mistreatment of others (e.g., spousal abuse). The alternative interpretation is that drinking is used to "forget" –both because thinking about one's condition is painful, and as a way of insulating future selves from today's depressing self–knowledge.

3 The psychological immune system

"Just as there was in his study a chest of drawers which he managed never to look at, which he went out of his way not to encounter when walking in or out, because in one drawer was held the chrysanthemum which she had given him on the first evening, ... so there was in him a place which he never let his mind approach, imposing on it, if necessary, the detour of a long reasoning... : it was there that lived the memories of happier days."

Marcel Proust, *Remembrance of Things Past.*

3.1 Awareness management as a psychological defense mechanism

The idea that people selectively recall past events and behaviors is an old one (e.g., Darwin's quotation). Freud, in particular, argued that people repress unwelcome memories and ascribe personal shortcomings to others in order to avoid threats to their ego. Much recent research has confirmed that people often engage in *beneffactance:* they remember successes and forget failures, and view themselves as instrumental for good but not bad outcomes.[18] Similarly, when they commit a bad act, they do not seem to draw the full inference about their personality. They reframe performance by trying to convince themselves that the act was not so bad ("he deserved it", "the damage was limited", "the evidence concerning the extent of damage is biased"); and they attribute the responsibility to others.[19]

The tension between "hot attributions" designed to maintain and enhance self-esteem and "cold attributions" following rational inference principles has also long been recognized. For

[18] See Zuckerman, M. (1979)
[19] See, e.g., Snyder (1985).

instance, Sartre (1953) argued that the individual must simultaneously know and not know the same information. Gur and Sackeim (1979) defined self-deception as a situation in which: a) the individual holds two contradictory beliefs; b) he is not aware of holding one of the beliefs; c) this lack of awareness is motivated.

This section builds on Gur and Sackeim's definition of self-deception, and shows how the SCM can provide an interface between the motivation and cognition aspects of self-deception. The basic idea is that the individual can, at a cost, affect the probability of remembering a given piece of information. Under complementarity and time inconsistency, there is an incentive to try to remember favorable signals and to forget unfavorable ones. This is the motivation part.[20] On the other hand, we maintain our rational inference postulate, so people realize that they have a selective memory or attention. This is the cognition part. Let us now state our hypotheses in more detail.

Hypothesis 4 (phenomenal self). *The individual is aware of only part of the full stock of his self-knowledge.*

It is a common theme of Freud (1938) and psychoanalysts that most of mental life is unconscious. Modern cognitive psychology also recognizes the fact that only part of the individual's accumulated stock of information is readily available for conscious, purposive processing and decision-making. Baumeister (1998) uses an analogy with the computer: the full stock of self-knowledge corresponds to the content of the hard disk, and the phenomenal self to the part which is displayed on the screen -or, we might say, present in "active memory".

The fact that cognitive events often cannot be observed by the person within whom they occur solves the apparent inconsistency associated with an individual simultaneously knowing and not knowing the information.[21] Furthermore, certain types of data are more likely to be remembered than others, and this process is subject to both internal and external influences: a) Self-relevant information is processed and encoded more thoroughly; similarly, recollection is stronger when behavior was the object of the individual's free choice rather than imposed by the situation or by others.[22] b) Information that is rehearsed often is better remembered (indeed, that is why we cram for an exam); conversely, if one is preoccupied or distracted when an event unfolds, one has greater difficulty in remembering the details of what happened.[23] c) Direct behavioral experience makes the information more accessible in memory, because later on recall is more likely to be

[20] For evidence on the selectivity of memory in relation to self–enhancement, see Korner (1950), Crary (1996), Mischel, Ebbesen and Zeiss (1976), Kunda and Sanitioso (1989), or Murray and Holmes (1993).
[21] See, e.g., Tedeschi (1986) for a discussion.
[22] E.g. Baumeister (1998).
[23] Schacter (1996).

activated by situational *cues*.[24]

These mechanisms seem to be at work in experiments where individuals who are asked by the experimenters to behave in a self-deprecating manner later report lower self-esteem than earlier, while persons who are asked to display self-enhancing behavior report higher self-esteem.[25] This may be due to the fact that they were led to rehearse unfavorable or favorable information about themselves, which may increase the probability of remembering this information later on. In addition, receiving positive feedback seems to trigger a (cue-based) "warm glow" effect, which automatically makes accessible to the individual other instances of himself in positive situations.[26]

Hypothesis 5 (awareness or memory management). *The individual can, at a cost, increase or decrease the probability of remembering an event and/or its interpretation.*

Hypothesis 5 posits that memory is endogenous. This endogeneity is what, in our model, gives rise to *selective memory*. Note that an equivalent (and complementary) interpretation is that of *selective attention* or *selective awareness*.

Let us provide some motivation for the idea that conscious memory is partly (but of course only partly!) influenced by motivation. Suppose that the individual wishes to remember good news and forget bad ones. He can rehearse good news; linger over praise or positive feedback;[27] choose to be more frequently in environments or with people who will remind him of (provide him with cues associated with) good past information; describe to others his brilliant performance as long as he stays within the domain of acceptance (he must stay within the realm of polite or sincere approval, and not trigger a negative reaction by others).[28]

Conversely, an individual can use the rehearsal process to minimize the impact of self-threatening information.[29] He can eschew situations and people who may remind him of bad news. He can tear up the picture of a former girlfriend, or, like the narrator in Proust's novel, avoid passing by a chest of drawers which contains cues to painful memories; he can work unusually hard to "forget" (really, not think about) a failed relationship or family problem, or even use drugs and alcohol.

The individual may also use a wide array of strategies to discount bad news. Interpersonal strategies may be useful to accomplish this. Fighting with someone who criticizes one's paper in a seminar may not be a bad strategy in this respect (of course, this has costs...). We can view

[24] Fazio and Zanna (1981).
[25] Jones et al (1981).
[26] Greenwald (1980).
[27] Baumeister (1998).
[28] See Rhodewalt (1986) for a discussion of such self-presentation strategies and their link with self- enhancement.
[29] Relatedly, Gilbert and Cooper (1985, 79) note that "we can expect [an author in a meeting] to spend more time considering the comments of the lone dissenter who praised the project (and confirmed his self-conception) than of the colleagues who disliked it, thus mercifully "softening" the cavalcade of criticism".

this strategy as creating a distraction that will impair accurate recollection of the details of the criticism. As Gilbert and Cooper (1985) argue, "social interaction is a fertile context for self-deception because its very complexity often acts as a "smoke screen", keeping the self-deceptive process from becoming obvious."

Another common way of discounting self-threatening news is to seek out information that derogates the informativeness of the initial data.[30] Gilovich (1991) argues that people resist the challenge of information not by ignoring it, but by subjecting it to particularly intense scrutiny. After being criticized in a seminar, a researcher will look for various reasons why the commenter has poor tastes or a vested interest in a competing theory or body of evidence, why the criticism is minor, and so forth.

It is important to note that *we need not literally assume that the individual can directly and mechanically suppress memories.* Our model is equally consistent with a Freudian view where memories get buried in the unconscious (with some probability of reappearance), and with the more recent cognitive psychology which holds that memory itself cannot be controlled, but emphasizes the different ways in which *awareness can be affected:* the choice of attention when the information accrues, the search for or avoidance of cues and the process of selective rehearsal afterwards, and again the choice of attention at the time the information is (voluntarily or accidentally) retrieved. While the underlying processes are very different, the end-result of these two views of motivated beliefs is formally equivalent: the individual has differential rates of recall or awareness, depending on how helpful or hurtful the information is to his self-esteem and general efficacy. We shall therefore use the terms "memory" and "awareness of past informations" interchangeably.

Hypothesis 5 suggests the following mathematical representation. Consider an information received at date 0 by the individual, and let $\lambda \in [0,1]$ denote the probability that he remembers (accesses) this information at date 1, when he decides whether to exert effort. Since there is a cost involved both in increasing λ and in decreasing it, one may define the natural probability of remembering $\lambda_N \in (0,1)$ as that which maximizes the date 0 flow payoff u_0. Increasing or decreasing λ involves a "memory cost" $M(\lambda)$, i.e., a reduction in u_0, with $M(\lambda_N) = 0$, $M'(\lambda) \leq 0$ for $\lambda < \lambda_N$ and $M'(\lambda) \geq 0$ for $\lambda > \lambda_N$.[31]

[30] Frey (1981).
[31] Although it might seem tautological, it is worth emphasizing that "forgetting" means that an actual loss of information (a coarsening of the informational partition) occurs. In particular, if at date 1 the individual "does not remember" a signal σ received at date 0, he also does not recall any other piece of information that is perfectly correlated with σ. Thus, when a different λ is chosen in response to different signals σ, if forgetting does occur the individual also does not recall which costs $M(\lambda)$ were incurred in the process. More generally, one could allow M to be a noisy function of λ, in which case remembering how hard one worked at repression (or how much one drank to forget) would still allow only an imperfect reconstruction of λ, and therefore of σ. Concretely, it must be that remembering that one tore up the picture of a former lover, or hid away a chrysanthemum, is not as painfully

It is worth emphasizing again that this mathematical formulation does not literally mean that the individual has direct control over his rate of recall, or over which memories to keep and which to forget. The natural rate of recall (or speed of memory decay) is determined by exogenous factors (salience, feedback from the environment, etc.). By using the various means discussed above the individual can, however, *indirectly manipulate* the probability that he will still be aware of the information when making a decision at date 1. One could even adhere to a minimalist version of the model where the individual can only improve his rate of recall (through rehearsal, record-making, etc.), but never lower it ($M(\lambda) = \infty$ for all $\lambda < \lambda_N$). All that matters is the potential for a differential rate of recall or awareness in response to desirable or undesirable informations.

Hypothesis 2' (awareness of selectivity of memory). *While the individual can manipulate his conscious self-knowledge, he is aware that incentives exist that result in selective memory.*

Hypothesis 2' is a special case of Hypothesis 2 in the awareness-management context. It posits that, even though individuals can manipulate their memory, they cannot systematically fool themselves. That is, they keep using rational inference and realize that what they may have forgotten are not random events.

Remark 1: Hypotheses 5 and 2' make our model reminiscent of Holmström's (1982, 1999) path-breaking work on career concerns. In Holmström's model, a worker as well as the labor market are unaware of his ability (θ in our model). The worker selects an unobservable effort and the labor market, observing his performance, updates its beliefs about his ability before making wage offers in the following period. In equilibrium, the worker exerts effort to try and fool the labor market (convince it of a high ability through a good performance); yet the labor market is not fooled (on average), as it rationally understands the worker's career objectives. Similarly, in our model, Self 0 may incur a cost to distort memory in order to fool Self 1, but Self 1 will rationally take this incentive into account.

Remark 2: While our memory model shares Hypothesis 4 with that of Mullainathan (1998), it differs crucially from it through Hypotheses 5 and 2'. First, it is a model where memory or awareness is purposefully distorted, while Mullainathan's paper (which has a different object) studies the consumption and savings implications of exogenous, naturally occurring, memory losses. Second, we assume rational (Bayesian) inference.

informative (about the fact that someone else was preferred, what one could/should have said or done, etc.) as actually staring at it. Otherwise, there would be no reason to do so. Similarly, knowing that one avoids certain people or places because they were witness to the failure of one's business, intellectual or marital venture is not as bad as being constantly reminded by them of just how badly one failed.

3.2 Self-deception through selective memory / awareness

"To break down the renewed assaults of my memory, my imagination effectively labored in the opposite direction."

Marcel Proust, *Remembrance of Things Past.*

The agent's ability is a random variable θ, about which he receives at date 0 a signal σ. As earlier, ability is persistent across periods, so that σ is informative about the probability of success of date 1 effort in generating date 2 payoffs.[32] The agent's Self 1 would therefore benefit from having this information, but if it is ego-threatening Self 0 may have an interest in suppressing or tampering with it, in order to preserve the effort incentives of Self 1. To make things as simple as possible, we assume that the signal received at date 0 can take only two values. With probability $1-q$ the agent receives bad news about his ability, $\sigma = L$, and with probability q he receives no news at all, $\sigma = \emptyset$. In other words, "no news is good news". Let:

$$\theta_L \equiv \mathrm{E}[\theta \,|\, \sigma = L] < \mathrm{E}[\theta \,|\, \sigma = \emptyset] \equiv \theta_H. \tag{13}$$

The recollection at date 1 of the news σ will be denoted $\tilde{\sigma}$. We assume that memories can be (partially) suppressed but not manufactured ex-nihilo, so that the issue of memory manipulation arises only following bad news.[33] In other words, $\sigma = \emptyset$ always leads to $\tilde{\sigma} = \emptyset$. A signal $\sigma = L$, on the other hand, may be forgotten due to natural forgetfulness or voluntary repression. Let λ denote the probability that bad news will be remembered accurately:

$$\lambda \equiv \Pr[\tilde{\sigma} = L \,|\, \sigma = L]. \tag{14}$$

We assume that the agent can, at a cost, increase or decrease this probability with respect to its "natural" value $\lambda_N \leq 1$. As discussed earlier, the recall probability may be increased by rehearsing the information, making non-manipulable written records of it, communicating it to outside parties, etc. It may be decreased by initially discounting the information, questioning the motives of the news-bearer, focusing attention on other (relevant or irrelevant) issues, through repression into the unconscious, drug or alcohol use, etc. As these examples show, the costs incurred in either case can be material (record-keeping, alienating others, physical effects of alcohol or substance abuse) or psychic (stress from repression, neurosis). We assume that choosing a recall probability λ involves a cost of $M(\lambda)$. We shall now analyze the equilibrium in several stages.

[32] The signal would generally also matter because it is informative about the success probability of date 0 effort in generating date 1 payoffs. This decision problem does not involve any issues of memory manipulation for self-esteem purpose: the information at date zero is given, and the agent simply chooses the best decision accordingly. We shall thus abstract from this decision problem.

[33] The topic of false memories and their "recovery" is left for further research.

1. Inference problem of Self 1. Faced with a memory $\hat{\sigma} \in \{L, \emptyset\}$, Self 1 must first assess its credibility. Given that memories cannot be invented, unfavorable ones are always credible (it is as if they were fully verifiable). When Self 1 does not recall any adverse signals, on the other hand, he must ask himself whether there were indeed no bad news at date 0, or whether they may have been lost due to either the natural fallibility of memory or to strategic forgetting. If Self 1 thinks that bad news are recalled with probability λ^*, he uses Bayes' rule to compute the *reliability* of a "no recollection" message as:

$$r^* \equiv \Pr[\sigma = \emptyset \mid \hat{\sigma} = \emptyset; \lambda^*] = \frac{q}{q + (1-q)(1-\lambda^*)}. \tag{15}$$

His degree of self-confidence is then:

$$\theta(r^*) \equiv r^* \theta_H + (1 - r^*)\theta_L. \tag{16}$$

A rational Self 1 will use in these computations the correct probability λ chosen by Self 0. For the moment we shall take his *belief* λ^* as a fixed parameter, and impose the equilibrium condition $\lambda^* = \lambda$ later on.

2. Decisions and payoffs. We normalize the payoff in case of success to $V = 1$, and assume that the cost of date 1 effort is drawn from an interval $[\underline{c}, \bar{c}]$, with probability distribution $\Phi(c)$ and density $\varphi(c) > 0$. We assume $\bar{c} > \beta\delta\theta_H > \beta\delta\theta_L > \underline{c}$, which means that there is always a positive probability of no effort, and a positive probability of effort, at date 1.[34]

Given a signal σ at date 0 and a memory $\hat{\sigma}$ at date 1, Selves 0 and 1 respectively assess the productivity of date 1 effort as $\mathsf{E}[\theta \mid \sigma]$ and $\mathsf{E}[\theta \mid \hat{\sigma}]$. Self 1 only works when the realization of the effort cost is $c < \beta\delta\mathsf{E}[\theta \mid \hat{\sigma}]$, so Self 0's payoff is:

$$\beta\delta \int_0^{\beta\delta\mathsf{E}[\theta \mid \hat{\sigma}]} (\delta\mathsf{E}[\theta \mid \sigma] - c)\, d\Phi(c). \tag{17}$$

To build intuition, suppose for a moment that Self 0 could freely and costlessly pick or manipulate Self 1's expectation, $\mathsf{E}[\theta \mid \hat{\sigma}]$. What beliefs would he choose for a naive Self 1? Maximizing (17), we find that Self 0 would like to set $\mathsf{E}[\theta \mid \hat{\sigma}]$ equal to $\mathsf{E}[\theta \mid \sigma]/\beta$. This makes clear how the time-consistency problem gives Self 0 an incentive to boost or maintain Self 1's self-confidence. The problem of course is that Self 1 is not so easily fooled.

[34] In Section 2 we took the distribution of θ to be continuous, and c was fixed. In this section c has a continuous distribution and θ can take only two values. The two formulations are actually isomorphic (even if the latter happens to be more convenient here): all that really matters is the distribution of $\delta\theta V/c$.

3. Costs and benefits of selective memory (or attention). Let us focus on the "bad news" case ($\sigma = L$) and compute the expected utility achieved by Self 0 when he follows an arbitrary recall strategy λ, given Self 1's beliefs λ^* –or equivalently, r^*. If a negative signal is truthfully interpreted and recalled, Self 0's expected welfare (gross of the cost of memory management) is:

$$U_T(\theta_L \mid r^*) = \beta\delta \int_0^{\beta\delta\theta_L} (\delta\theta_L - c)\, d\Phi(c), \tag{18}$$

where the subscript T stands for "truth". This expression is independent of r^*, so we shall denote it as just $U_T(\theta_L)$. If the bad news are successfully censored or repressed, on the other hand, Self 0 achieves:

$$U_C(\theta_L \mid r^*) = \beta\delta \int_0^{\beta\delta\theta(r^*)} (\delta\theta_L - c)\, d\Phi(c), \tag{19}$$

where the subscript C stands for "censored". The net gain or loss from *self-deception* is thus:

$$U_C(\theta_L \mid r^*) - U_T(\theta_L) = \beta\delta \int_{\beta\delta\theta_L}^{\beta\delta\theta(r^*)} (\delta\theta_L - c)\, d\Phi(c). \tag{20}$$

Hiding from Self 1 the information that $\theta = \theta_L$ leads him to exert effort in additional states of the word, namely those where $\beta\delta\theta_L < c < \beta\delta\theta(r^*)$. Like the self-handicapping behavior discussed earlier, this generally involves both costs and benefits. If r^* is high enough that $\beta\theta(r^*) > \theta_L$ (which requires $\beta\theta_H > \theta_L$), then:

$$U_C(\theta_L \mid r^*) - U_T(\theta_L) = \beta\delta \left(\int_{\beta\delta\theta_L}^{\delta\theta_L} (\delta\theta_L - c)\, d\Phi(c) - \int_{\delta\theta_L}^{\beta\delta\theta(r^*)} (c - \delta\theta_L)\, d\Phi(c) \right). \tag{21}$$

The first integral is decreasing in β, becoming zero at $\beta = 1$: it represents the gains from *confidence-building*, which alleviates Self 1's procrastination problem. The second integral is positive and increasing in β : it reflects the loss from *overconfidence*, which causes Self 1 to work in states of the world where his net productivity is actually so low that even Self 0 would prefer that he not exert effort. Note that *the loss from overconfidence is most likely to arise, the more reliable Self 1 considers the memory process to be*, i.e. the larger r^*. Conversely, if r^* is so low as to have $\beta\theta(r^*) < \theta_L$ (which requires $\beta\theta(q) < \theta_L$) the overconfidence effect disappears entirely, and we see from (20) that $U_C(\theta_L \mid r^*) > U_T(\theta_L)$.

4. Strategic memory or awareness management. Faced with as signal $\sigma = L$ which is hurtful to his self-esteem, Self 0 chooses the recall probability λ so as to solve:

$$\max_{\lambda} \{\lambda U_T(\theta_L) + (1-\lambda)U_C(\theta_L \mid r^*) - M(\lambda)\}. \tag{22}$$

Given the convexity of $M(\lambda)$, the optimum is uniquely determined (given r^*) by the first-order condition:

$$\begin{cases} \text{if } U_C(\theta_L \mid r^*) - U_T(\theta_L) + M'(0) \geq 0, \text{ then } \lambda = 0; \\ \text{if } U_C(\theta_L \mid r^*) - U_T(\theta_L) + M'(1) \leq 0, \text{ then } \lambda = 1; \\ \text{if } U_C(\theta_L \mid r^*) - U_T(\theta_L) \in (-M'(1), -M'(0)), \text{ then } \lambda = \Lambda(r^*, \beta), \end{cases}$$

where $\Lambda(r^*, \beta) \in (0,1)$ is the unique solution to:

$$\beta\delta \int_{\beta\delta\theta_L}^{\beta\delta\theta(r^*)} (\delta\theta_L - c)\, d\Phi(c) + M'(\lambda) = 0. \tag{23}$$

Finally, the Bayesian rationality of Self 1 means that he is aware of Self 0's choosing the recall strategy λ opportunistically according to (22), and uses this optimal λ in his assessment of the reliability of memories (or lack thereof).

Definition 2 *A Perfect Bayesian Equilibrium (PBE) of the memory game is a pair* $(\lambda^*, r^*) \in [0,1] \times [q,1]$ *such that:*

i) The recall strategy of Self 0 is optimal, given Self 1's assessment of the reliability of memory:

$$\lambda^* \in \arg\max_{\lambda} \{\lambda U_T(\theta_L) + (1-\lambda) U_C(\theta_L \mid r^*) - M(\lambda)\}$$

ii) Self 1 assesses the reliability of memories using Bayes' rule and Self 0's recall strategy:

$$r^* = \frac{q}{q + (1-q)(1-\lambda^*)}.$$

In the remainder of this section we shall investigate two main issues:

1. Nature and multiplicity of equilibria. What modes of self-esteem management are sustainable (e.g., from "systematic denial" to "complete self-acceptance"), depending on a person's characteristics such as his time-discounting profile or cost of memory manipulation? Can the same person, or otherwise similar people, be "trapped" in different modes of behavior?

2. Welfare analysis. Is a more active self-esteem maintenance strategy always beneficial, or can it end up being self -defeating? Would a person rather be free to manage their self-confidence and memories as they see fit, or prefer *a priori* to find mechanisms (friends, mates, environments, occupations, etc.) which ensure that they will always be confronted with the truth about themselves, no matter how unpleasant it turns out to be? Formally, we shall study how different equilibria are ranked in terms of ex-ante welfare, and how they compare with a strategy of always truthful memories.

Because PBEs are related to the solutions $r^* \in [q,1]$ to the nonlinear fixed-point equation

$$\psi(r,\beta) \equiv \beta\delta \int_{\beta\delta\theta_L}^{\beta\delta(r\theta_H + (1-r)\theta_L)} (\delta\theta_L - c)\, d\Phi(c) + M'\left(\frac{1-q/r}{1-q}\right) = 0 \tag{24}$$

that is obtained by substituting (15) into (23), we will use a sequence of simpler cases to demonstrate the main points which emerge from our framework. Note, for further reference, that $\psi(r,\beta)$ represents Self 0's (net) *marginal incentive to forget.*

3.3 Costless memory or awareness management

We shall solve here the memory game in the case where the manipulation of memory is costless, $M \equiv 0$. While it does not allow us to address the psychological costs of repressed memories (as opposed to their informational ones), this case already yields several key insights and results, and is very tractable. Note also that while we shall assume that λ can be freely varied between 0 and 1, the results would be identical if it were constrained to lie in some interval $[\underline{\lambda}, \bar{\lambda}]$. With $\underline{\lambda} > 0$ one can never forget (or avoid undesired cues) for sure.[35] Once again, all that matters for our theory is the potential for differential probabilities of recall or awareness, depending on whether the information is harmful or beneficial to one's self-esteem and efficacy.

Proposition 2 *Assume costless memory/awareness management* ($M \equiv 0$). *For low degrees of time inconsistency there is minimum repression, for high degrees there is maximum repression. For intermediate degrees of time inconsistency there are three equilibria: minimum repression, maximum repression, and a mixed strategy or partially repressive equilibrium.*

More formally, there exist $\underline{\beta}$ *and* $\bar{\beta}$ *in* $(0,1)$, $\underline{\beta} < \bar{\beta}$, *such that:*

i) For all $\beta > \bar{\beta}$, *the unique PBE corresponds to* $\lambda^* = 1$;

ii) For all $\beta < \underline{\beta}$, *the unique PBE corresponds to* $\lambda^* = 0$;

iii) For all β *in* $[\underline{\beta}, \bar{\beta}]$, *there are three equilibria, corresponding to* $\lambda^* \in \{0, \Lambda(\beta), 1\}$, *where* $\Lambda(\beta)$ *decreases continuously from* 1 *to* 0 *as* β *rises from* $\underline{\beta}$ *to* $\bar{\beta}$.

These results are illustrated on Figure 1. The intuition is simple and can grasped from (21), where we saw that the confidence-building effect increases with time inconsistency $1-\beta$, while the overconfidence effect decreases with it. When β is high enough, overconfidence is the dominant concern, and adverse signals are systematically transmitted. Conversely, for low enough values of β the confidence-building effect dominates, so ego-threatening signals are systematically

[35] It could even be that the natural rate of recall is $\lambda_N = \underline{\lambda}$. In that most restrictive case, the individual can only slow down the natural rate of memory decay (increase λ above λ_N), but not speed it up.

forgotten. For intermediate values of β, finally, the two effects are relevant and there may be multiple equilibria, including one where memory is partially selective.

Note that, in the type of second equilibrium, none of Self 0's information is ever transmitted to Self 1: with $\lambda^* = 0$, having "no recollection of anything bad" is completely uninformative as to whether or not something bad occurred ($r^* = q$). In the language of communication games, this is a "babbling equilibrium". The mechanism at work here is quite different from the *ex ante* suppression of information considered earlier when analyzing self-handicapping, or in Carrillo and Mariotti (1997). Self 0 *does not want* to suppress good news, only bad news; but in doing the latter *he cannot help* but also do the former. As we shall see later on, this may end up doing him more harm than good, whereas the usual "strategic ignorance" is only chosen when it improves ex-ante welfare.

What makes all three equilibria self-fulfilling is precisely the *introspection* or "metacognition" of the Bayesian individual, who understands that his cognitive process (in this instance, his memory) is subject to opportunistic distortions. The higher the degree of censoring by Self 0, the more Self 1 discounts the "no bad news to report" recollection, and therefore the lower the risk that he will be overconfident. As a result, the greater is Self 0's incentive to censor. Conversely, if Self 0 faithfully records all news in memory, Self 1 is more likely to be overconfident when he cannot recall any bad signals, and this incites Self 0 to be truthful. Finally, consider the mixed-strategy equilibrium. Since a higher β lowers the payoff to self-esteem maintenance, this must be offset by a smaller loss from overconfidence –and thus by a lower reliability of the "no bad news" report– in order to keep Self 0 just indifferent between censoring and not censoring. This is why $\Lambda(\beta)$ and $R(\beta)$ are decreasing in β : in this equilibrium, memory is more selective or biased, the less severe is the time-inconsistency problem.

The last observation to be drawn from Figure 1 is that, as β rises from 0 to 1, there is necessarily (i.e., for any equilibrium selection) at least one point where λ^* must have an upward discontinuity; it may also, but need not, have downward discontinuities. Thus, very small differences in the (psychic or material) costs of memory management, repression, etc., can imply large changes in the selectivity of memory, hence in the variability of self-confidence, and ultimately in performance.

3.4 Costly memory management

"Denial ain't just a river in Egypt." Mark Twain.

In this section we use specific functional forms to study the problem set up in Section 3.2. The memory cost function is:

$$M(\lambda) = a(1 - \ln \lambda) + b(1 - \ln(1 - \lambda)), \tag{25}$$

with $a > 0$ and $b \geq 0$. It minimized at the "natural" recall rate $\lambda_N = a/(a+b)$, and involves infinite costs of complete repression. When $b > 0$ perfect recall is also prohibitively costly, and M is U–shaped. As to the distribution of effort costs, we take it to be uniform, $\varphi(c) = 1/\bar{c}$ on $[0, \bar{c}]$, with $\bar{c} > \beta\delta\theta_H$. With these assumptions, the *incentive to forget* defined in (24) becomes:

$$\psi(r,\beta) = \left(\frac{\beta\delta}{\bar{c}}\right)\left(\delta\theta_L\left[\beta\delta\theta(r) - \beta\delta\theta_L\right] - \frac{1}{2}\left[(\beta\delta\theta(r))^2 - (\beta\delta\theta_L)^2\right]\right) + b\left(\frac{1-q}{q/r-q}\right) - a\left(\frac{1-q}{1-q/r}\right)$$

or

$$\psi(r,\beta) = r\,(\Delta\theta)\left(\frac{\beta^2\delta^3}{\bar{c}}\right)\left((1-\beta)\theta_L - \frac{\beta r}{2}\,(\Delta\theta)\right) - r\left(\frac{1-q}{q}\right)\left(\frac{(a+b)q-(aq+b)r}{(1-r)\,(r-q)}\right). \tag{26}$$

where $\Delta\theta \equiv \theta_H - \theta_L$. Inspection of (26) makes clear that the sign of $\psi(r,\beta)$ is that of a third degree polynomial in r, with $\lim_{r\to q}\psi(q,\beta) = -\infty$ as long as $a > 0$, and $\lim_{r\to 1}\psi(q,\beta) = +\infty$ as long as $b > 0$. Therefore, there are either one or three solutions to $\psi(r,\beta) = 0$ for a given β, meaning either one or three equilibria, as in the previous section. One could easily study this polynomial numerically, or find sufficient conditions for its having three roots in $[q,1]$. But since we also want to study how the equilibria vary with key parameters like β and a, we shall focus on the simpler case where *recall is costless* but *repression is costly.*

These results are formally described in the next proposition, but perhaps more easily grasped from Figures 2.1 to 2.3.

Proposition 3 *Assume that φ is uniform on $[0,\bar{c}]$ with $\bar{c} > \beta\delta\theta_H$, and that $b = 0$. A higher degree of time inconsistency or a lower cost of repression increases the scope for memory manipulation by generating partially repressive (mixed strategy) equilibria, and possibly making perfect recall unsustainable.*

More formally, there exist $\beta_1 < \beta_2 < \beta_3$ and continuous functions $\lambda_1(a)$, $\lambda_2(a)$, respectively decreasing and increasing in the repression cost parameter a, such that:

i) For all $\beta \geq \beta_3$, the unique PBE corresponds to $\lambda^* = 1$;

ii) For $\beta_2 \leq \beta < \beta_3$ there exists $\bar{a} > 0$ such that the equilibrium set corresponds to $\lambda^* = 1$ for $a > \bar{a}$, and to $\lambda \in \{\lambda_1(a), \lambda_2(a), 1\}$ for $a \leq \bar{a}$ (cf. Figure 2.1).

iii) For $\beta_1 \leq \beta < \beta_2$ there exists $\bar{a} > \underline{a} > 0$ such that the equilibrium set corresponds to $\lambda = 1$ for $a > \bar{a}$, to $\lambda \in \{\lambda_1(a), \lambda_2(a), 1\}$ for $\underline{a} \leq a \leq \bar{a}$, and to $\lambda = \lambda_1(a)$ for $a < \underline{a}$ (cf. Figure 2.2).

iv) For $\beta < \beta_1$ there exists $\underline{a} > 0$ such that the equilibrium set corresponds to $\lambda^* = 1$ for $a \geq \underline{a}$, and to $\lambda = \lambda_1(a)$ for $a < \underline{a}$ (cf. Figure 2.3).

The effects of β are thus essentially similar to those obtained earlier. The effects of a are also intuitive, except that lower costs of memory manipulation or repression need not always lead to

a more active memory-management strategy: the function λ_2 is decreasing in a. Finally, small changes in a can induce large changes in behavior and the variability of self-esteem.[36]

3.5 Welfare analysis of self–deception

"The art of being wise is the art of knowing what to overlook."

William James, *Principles of Psychology (1890).*

"There is nothing worse than self-deception –when the deceiver is at home and always with you."

Plato (quoted by Mele 1997).

Is a person ultimately better off in an equilibrium with a strategy of active self-esteem maintenance and "positive thinking" ($\lambda^* < 1$), or when he always faces the truth? Like Plato and William James, psychologists are divided between these two conflicting views of self-deception. On one side are those who endorse and actively promote the self-efficacy / self-esteem movement (e.g., Bandura 1977, Seligman 1990), pointing to studies which tend to show that a moderate dose of "positive illusions" has significant affective and functional benefits. On the other side are skeptics and outright critics (e.g., Baumeister 1998, Swann 1996), who see instead a lack of convincing evidence, and point to the dangers of overconfidence as well as the loss of standards which results when negative feedback is systematically withheld or discounted in the name of self-esteem preservation. Our model will provide insights into the reasons for this ambiguity.

Consider an equilibrium with recall probability $\lambda^* \leq 1$ and associated credibility r^* (via (15)). With probability $1-q$, Self 0 receives bad news, which he then forgets with probability $1-\lambda^*$; the resulting expected payoff is $\lambda^* U_T(\theta_L) + (1-\lambda^*)U_C(\theta_L \mid r^*) - M(\lambda^*)$. With probability q the news are good, which means that no adverse signal is received. The problem is that the credibility of a "no bad news" memory in the eyes of Self 1 may be quite low, so that he will not exert much effort even when it is actually optimal to do so. Thus, the payoff to Self 0 following genuinely "good news" is only

$$U_T(\theta_H \mid r^*) = \beta\delta \int_0^{\beta\delta\theta(r^*)} (\delta\theta_H - c)d\Phi(c), \tag{27}$$

which is clearly less than $U_T(\theta_H \mid 1)$ whenever cost realizations between $\theta(r^*)$ and $\theta(1)$ have positive probability. In that case there is a loss from *self-distrust* or *self–doubt*, compared to a situation where Self 0 always truthfully records all events into memory. More generally, the

[36] It might also be interesting to note that the model's specification with a uniform distribution of c on $[0, \bar{c}]$ and $\bar{c} > \beta\delta\theta_H$ is formally equivalent to one where c is fixed (say, $c \equiv 1$) but effort is a continuous decision, with net discounted payoff $\beta\delta\theta e - e^2/2$ for Self 1 and $\beta\delta\left(\delta\theta e - e^2/2\right)$ for Self 0.

agent's ex-ante welfare in equilibrium equals

$$\mathcal{W}(\lambda^*, r^*) \equiv qU_T(\theta_H \mid r^*) + (1-q)\left[\lambda^* U_T(\theta_L) + (1-\lambda^*)U_C(\theta_L \mid r^*) - M(\lambda^*)\right]. \tag{28}$$

Let us now assume that truth (perfect recall) is also an equilibrium strategy, with cost $M(1)$; as we shall see, a very similar analysis applies if $\lambda^* = 1$ is achieved by using some *a priori* commitment mechanism (chosen before σ is observed). Denoting the difference in welfare with this benchmark case as $\Delta W(\lambda^*, r^*) \equiv \mathcal{W}(\lambda^*, r^*) - \mathcal{W}(1,1)$, we have:

$$\begin{aligned}\Delta W(\lambda^*, r^*) &= (1-q)\left[\left((1-\lambda^*)\left(U_C(\theta_L \mid r^*) - U_T(\theta_L)\right) - M(\lambda^*) + M(1)\right)\right] \\ &\quad -q\left[U_T(\theta_H \mid 1) - U_T(\theta_H \mid r^*)\right],\end{aligned}$$

or, finally:

$$\begin{aligned}\Delta W(\lambda^*, r^*) &= (1-q)\left((1-\lambda^*)\int_{\beta\delta\theta_L}^{\beta\delta\theta(r^*)} (\delta\theta_L - c)d\Phi(c) - M(\lambda^*) + M(1)\right) \\ &\quad -q\int_{\beta\delta\theta(r^*)}^{\beta\delta\theta_H} (\delta\theta_H - c)d\Phi(c).\end{aligned} \tag{29}$$

The first expression describes the net *gain from forgetting bad news*, the second one the *loss from disbelieving good news*. A few general results can be immediately observed.

First, *if memory manipulation is costless* ($M = 0$), then a *partial recall* (mixed strategy) equilibrium, when it exists, *cannot be better than perfect recall*. Indeed, in such an equilibrium the gain from hiding bad news is zero ($U_C(\theta_L \mid r^*) = U_T(\theta_L)$), because the self-enhancement and overconfidence effects just cancel out. The cost from self-distrust, on the other hand, is always present.

When repression is costly, this reasoning does not apply any more, as the term in the first set of brackets becomes $M(1) - M(\lambda^*) - (1-\lambda^*)M'(\lambda^*) > 0$, by the convexity of M. The ranking of the mixed strategy and perfect recall equilibria, when both exist, is thus a priori ambiguous. Similarly, when systematic repression or denial ($\lambda^* = 0$) is an equilibrium, it generates a positive "surplus" in the event of bad news: $U_C(\theta_L \mid q) - U_T(\theta_L) > M'(0) \geq 0$. The issue remains, in both cases, of how this gain compares to the loss from self-distrust. We shall show below that the balance can go either way, so that:

1) In some situations, an active strategy of self-esteem maintenance, selective memory, "looking on the bright side', avoiding "negative" thoughts and people, etc., as advocated in numerous "self-help" books, may actually pay off.[37]

[37] A brief search for "self-esteem" on the web gives an idea of this huge industry. There are numerous sites

2) In other situations, one can only lose by playing such games with oneself , and it would be better to "be honest with yourself", and "accept who you are".

The second case is also interesting because it involves *two degrees of lack of commitment:* it is because the agent cannot commit to working in period 1 that his inability to commit not to tamper with memory in period 0 becomes an issue, which may end up hurting him more than if he had simply resigned himself to the original time-consistency problem.

As suggested by the discussion of (29), the key intuition as to whether Case 1 or Case 2 applies is the likelihood of cost realizations sufficiently high to discourage effort in the absence of adverse recollections, $\tilde{\sigma} = \emptyset$. When such events are infrequent the self-distrust effect is small or even absent and, on average, self-deception pays off. When they are relatively common, the reverse is true.

Proposition 4 *Let $M \equiv 0$. If the function $\Gamma(Z,\beta) \equiv \int_0^Z (Z - \beta c) d\Phi(c)$ is concave in Z, then ex ante welfare is higher if all bad news are censored from memory than if they are always recalled. The reverse is true if Γ is convex in Z. The function Γ is concave in Z if and only if the cost density $\varphi(c)$ decreases fast enough:*

$$-\frac{\partial \ln \varphi(c)}{\partial \ln c} > \frac{2-\beta}{1-\beta}, \quad \textit{for all } c \in [0, \bar{c}].$$

It is convex if the inequality is reversed.

The proposition also shows that, for a given cost distribution φ, self-deception is more likely to be beneficial for a less time–consistent individual. This was far from obvious a priori, since both the gain and the loss in (29) decrease with β: in equilibrium, memory manipulation tends to alleviate procrastination when $\sigma = L$, but worsen it when $\sigma = \emptyset$. The underlying intuition is relatively simple, however. The net loss across states from a "hear no evil–see no evil" strategy $\lambda^* = 0$, namely $-\Delta W(q,0)$, is simply the *ex-ante* value of information (always recalling the true σ, rather than having only the uninformed prior $\Pr[\sigma = \emptyset] = q$). As is well known (and was seen in Section 2.2 for the simpler case where c is deterministic), it is only when time–consistency is strong enough that this value can be negative.

Let us now consider a few examples of welfare rankings.

devoted to the subject, and hundreds of books with titles such as: *"How to Raise Your Self-Esteem", "31 Days to High Self-Esteem : How to Change Your Life So You Have Joy, Bliss & Abundance", "365 Ways to Build Your Child's Self–Esteem", "501 Ways to Boost Your Child's Self–Esteem", "611 Ways to Boost Your Self Esteem : Accept Your Love Handles and Everything About Yourself", "ABC I Like Me"*, etc.

a) With a uniform density on $[0,\bar{c}]$ ($\bar{c} > \beta\delta\theta_H$), self-deception is *always hurtful* compared to truth-telling. This applies whether both $\lambda^* = 0$ and $\lambda = 1$ are in the equilibrium set, or only one of them. Proposition 2 and Figure 1 show when each case applies, depending on the value of β.

b) Conversely, self-deception is *always useful* when $\varphi(c) = \gamma c^{-n}$ on $[\underline{c}, +\infty)$, with $0 < \underline{c} < \beta\delta\theta_L$, γ chosen so that the density sums to one, and $n > (2-\beta)/(1-\beta)$. In this case it can also be shown that $\lambda^* = 0$ is the only equilibrium.

c) Finally, we provide in the appendix a simple example where both $\lambda^* = 0$ and $\lambda^* = 1$ are equilibria (simultaneously), and where *either one* -depending on parameter values- may lead to higher ex-ante welfare.

We have thus far interpreted the "always face the truth" strategy as an equilibrium, if sustainable alongside with λ^*. Alternatively, it could result from some initial commitment of the type discussed earlier (chosen before σ is observed), which amounts to *making oneself face steeper costs of self-deception* (increasing $M(\lambda)$ for $\lambda < \lambda_N$, and/or increasing it for $\lambda > \lambda_N$).[38] This reinterpretation requires minor modifications to (29), but the main conclusions remain unaltered.[39]

Interestingly, the multiplicity and welfare ranking of intrapersonal equilibria indeed provides a role for benevolent outside parties, such as parents or therapists, to help an individual escape the "self-traps"(Swann 1996) in which he might be stuck: depressive state of low self-esteem, chronic blindness to his own failings, etc. They can make him aware that a better personal equilibrium is feasible, and teach him how to coordinate on it by following certain simple rules with respect the gathering, interpretation, and encoding into memory of self-relevant information. They may also offer a form of informational commitment, serving as the repositories of facts and feelings which the individual realizes that he or she has an incentive to forget ("let's talk about that incident with your mother again"). More generally, they allow him to alter the "awareness/repression" technology $M(\lambda)$, whether through their own feedback and questioning, or by teaching him certain cue-management techniques. Indeed, much of modern cognitive therapy (as well as early education) aims at changing people's beliefs about themselves through selective recollection and rehearsal of events, self-serving attributions about success and adversity, and similar paths to "learned optimism" (to use the phrase coined by Seligman 1990).

[38] For instance, an individual with $\beta < \bar{\beta}$ in Proposition 2 may be worse off when memory management is free ($M = 0$) than when repression is prohibitively costly ($M(\lambda) = +\infty$ for all $\lambda \neq 1$). For instance, case (a) above shows that such is always the case when the cost distribution $\varphi(c)$ is uniform.

[39] Term $M(1)$ in the (28)–(29) is simply replaced by $\beta^{-1}\delta^{-\tau}\tilde{M}/(1-q) + m(1)$, where $\tilde{M}$ is the up–front cost of the commitment mechanism, $-\tau < 0$ is the period when the commitment was made, and $m(1) \geq 0$ is the cost of perfect recall faced at $t = 0$ as a result of this decision (whereas $m(\lambda) = +\infty$ for all $\lambda < 1$).

4 Beliefs and Make-Beliefs [40]

Our strategy so far has been to depart as little as possible from the view that people are fully rational decision-makers and information processors (they simply have non-exponential discount rates), and show that one can still explain a wide range of apparently "irrational" behaviors such as self-handicapping, selective awareness or memory, or defensive pessimism (see Section 5). Of course, the Bayesian assumption is not uncontroversial among psychologists.

One might object for instance that experiments consistently suggest that people display biases in their processing of information, and that surveys as well as daily observation consistently suggest that they overestimate their abilities and other desirable traits. Well-educated, reflective individuals seem to be no exception since, as Gilovich (1991) relates, *"a survey of college professors found that 94% thought they were better than their average colleague"*.

Without denying the validity of the above-mentioned type of evidence, we would first like to emphasize that it should be interpreted with caution. Answers to surveys or experimental questionnaires may reflect self-presentation motives (for the benefits of the interviewer), or selective memory rehearsal strategies (for the individual's own benefit, as predicted by our model). Since the cost of lying is typically zero, even minute potential benefits (psychic or tangible, internal or external), can lead the individual to overstate his true self-perceptions. Secondly, for every person who is "overconfident" about how great they are (professionally, intellectually, socially, maritally), another one may be found who is underconfident, depressed, paralyzed by guilt and self-doubt, but unlikely to acknowledge this to anyone except his closest confident, counselor, or therapist. These could even be the same people at different points in time.

Second, we will show in Section 4.1 that Bayes' rules is in fact quite consistent with most people overestimating their ability, whether in absolute terms or relative to others. The only constraint is that there be no bias on average in the population, but this precludes neither a positive median bias, nor self-deception strategies having aggregate real effects.

Having said this, we recognize that perfect inference probably presumes too much rationality, and that it may be more realistic to view agents as *imperfect Bayesians* who partially succeed in ignoring the process by which they arrived at their current beliefs. At the other extreme, taking beliefs as completely naive would be even more unrealistic. As illustrated by the quotations from Nietzsche and Darwin at the beginning of the paper, if a person consistently destroys, represses, or manages not to think about negative news, he will likely become aware that he has this systematic tendency, and realize that the absence of adverse evidence or recollections should not be taken at face value. This *introspection* is the fundamental trait of the human mind which the Bayesian assumption captures in our model. Without it, self-delusion would be very easy and,

[40] The terminology of "beliefs and make–beliefs" is borrowed from Ainslie (1999).

when practiced, always optimal (ex-ante). With even *some* of this metacognition, self-deception becomes a much more subtle and complex endeavor. We will thus show in Section 4.2 that the results established so far are robust to allowing for imperfections and biases in inference, as long as the individual's self-beliefs are sufficiently reactive to his actual pattern of behavior.

4.1 Optimistic and Pessimistic Biases

Continuing to work with the awareness-management version of the SCM, let us now compare the cross-sectional distributions of true and self-perceived abilities. In a large population, a proportion $1-q$ of individuals are of low ability, $\theta = \theta_L$ (more generally, low expected ability), having received a negative signal, $\sigma = L$. The remaining q, having received $\sigma = \emptyset$, have high (expected) ability, $\theta = \theta_H$. Average ability is $q\theta_H + (1-q)\theta_L = \theta(q)$; we shall assume that $q < 1/2$, so that median ability is θ_L.

Consider now the distribution of self-evaluations. Suppose for simplicity that, when faced with ego-threatening information ($\sigma = L$), everyone uses the same censoring probability $\lambda^* \in (0,1)$.[41] As before, let r^* denote the corresponding reliability of memory, given by (15). Thus, when individuals make decisions at date 1,

– a fraction $(1-q)(1-\lambda^*)$ *overestimate* their ability by $\theta(r^*) - \theta_L = r^*(\theta_H - \theta_L)$;

– a fraction q *underestimate* it by $\theta_H - \theta(r^*) = (1-r^*)(\theta_H - \theta_L)$.

If the costs of repression/forgetting are low enough (e.g., a small a on Figure 2.3), one can easily have $(1-q)(1-\lambda^*) > 1/2$, perhaps even $(1-q)(1-\lambda^*) \lessapprox 1$. Thus most people believe themselves to be *more able than they actually are, more able than average, and more able than the majority of individuals.*[42] Adding those who had truly received the signal $\sigma = \emptyset$, the fraction of the population who think they are better than average is even larger, namely $1 - \lambda^*(1-q)$. The remaining minority think, correctly, that they are worse than average; as a result they have low motivation, undertake little, and achieve even less. They fit the experimental findings of depressed people as "sadder but wiser" realists, compared to their non-depressed counterparts who are much more likely to exhibit self-serving delusions.[43]

A seen above, Bayes' law does not constrain the skewness in the distribution of biases (this was first pointed out by Carrillo and Mariotti (1997), and is a feature which our model also shares with those of Brocas and Carrillo (1999) and Köszegi (1999). It only requires that the *average bias* across the $(1-q)(1-\lambda^*)$ optimists and the q pessimists be zero: and indeed,

[41] Either this is the unique equilibrium, as in Figure 2.3, or else we focus on a symmetric situation for simplicity.

[42] See e.g., Taylor and Brown (1988) and the other references mentioned following Hypothesis 2. Note that the above statement (like most experimental data) is about how the agent perceives his rank in the distribution of true abilities –not in the distribution of self–assessments which, as a Bayesian, he realizes are generally overoptimistic.

[43] Alloy and Abrahamson (1979).

$(1-q)(1-\lambda^*)r^* - q(1-r^*) = 0$ by (15). In other words, Bayesian rationality only imposes a tradeoff between the relative proportions of overconfident versus underconfident agents in the population, and their respective degrees of over- or under-confidence. Note, however, that a zero average bias in no way precludes self-esteem maintenance strategies from having *aggregate economic effects*. Clearly in our model, they do affect aggregate effort, output and welfare, as none of these is a linear function of perceived ability.

4.2 To Bayes or Not to Bayes?

Let us now relax the assumption of exact Bayesian updating, and allow the agent at date 1 to remain less than fully aware of the fact that, at date 0, any negative signal $\sigma = L$ would have been forgotten with probability λ. In other words, the agent now fails to remember not just *what* he forgets, but also *that* he forgets. Self 1's assessment of the reliability of a recollection $\hat{\sigma} = \emptyset$ is thus modified to be

$$r_\pi(\lambda) \equiv \Pr[\sigma = \emptyset \mid \hat{\sigma} = \emptyset; \lambda] = \frac{q}{q + \pi(1-q)(1-\lambda)}, \tag{30}$$

where λ is the actual recall strategy and $\pi \in [0,1]$ parametrizes the extent of Bayesian rationality. For $\pi = 0$ the agent is completely naive, and takes all recollections and perceptions at full face value. At the other extreme, $\pi = 1$ corresponds to (15), which is full rationality. Figure 3 depicts the inference relationship for all cases in-between: a lower π raises the perceived reliability r_π (or the perceived truthfulness $1 - \pi(1-\lambda)$), and makes it less responsive to the memory strategy which the agent actually uses.

Note first that, for a given λ, the value of π makes no difference to the proportions of agents who overestimate, underestimate, or correctly estimate their ability; as seen above, these are independent of how r^* is determined. Where imperfect inference does matters is in the *size* of these biases: by increasing r_π, a lower π weakens what we termed the self-doubt effect, therefore magnifying the overconfidence $r_\pi(\theta_H - \theta_L)$ of agents who have repressed a negative signal, and simultaneously reducing the underconfidence $(1-r_\pi)(\theta_H - \theta_L)$ of those who discount the veracity of good news which, in their case, are in fact genuine. As a result, the average bias in the population (or for an individual over long enough periods) is now $(1-\pi)(1-\lambda)(1-q)r_\pi > 0$, and decreasing in π (given λ).

Let us now turn the determination of the equilibrium set with this more naive updating mechanism. To simplify the exposition, we shall focus without loss of generality on the case of costless awareness management ($M \equiv 0$). Figure 3 then depicts, as a bold step-function, Self 0's optimal recall probability λ, or *best response*, as a function of Self 1's belief r.[44] It is clear from

[44] The picture is drawn for parameters such that both the confidence-building and the overconfidence concerns

the graphs that if Self 1 always takes the absence of bad news at face value ($\pi = 0$), the *perfect recall* strategy $\lambda = 1$ is the unique equilibrium.[45] For a sophisticated Bayesian, on the other hand ($\pi = 1$), Figure 3 clearly illustrates the three equilibria of Proposition 1. As the degree of bias in Self 1's inference increases, the $\lambda = r_\pi$ curves rotates and flattens as described above. The three equilibria at first remain unchanged, except that the recall probability λ in the mixed equilibrium gradually declines. Beyond the level of π such that $r_\pi(0) = R(\beta)$, however, only the truthful equilibrium survives, as in the case of a naive Self 1.

The lesson from this simple extension of the model is that, as long as people are reasonably (even if not fully) objective in their inference, the results of multiplicity (and ex-ante welfare rankings) of personal equilibria which distinguish the Bayesian model from the naive one go through. Together with the results from the previous subsection, this leads us to conclude that whereas the explanatory power gained by departing from Bayesian rationality is somewhat limited (perception biases need no longer sum to zero), much can be lost if the departure is too drastic: without sufficient introspection, one can not account for "self-traps".

5 Variants and Extensions

5.1 Defensive Pessimism

While people are most often concerned with boosting their self-esteem, there are also instances where they seek to minimize their achievements, or convince themselves that the task at hand will be difficult rather than easy. For instance, a student studying for exams may discount his success on previous ones as attributable to luck or lack of difficulty. A young researcher or lawyer may understate the value of his prior achievements, as compared to what will be required to obtain tenure or partnership. A dieting person who lost a moderate amount of weight may decide that he "looks fatter than ever", no matter what others or the scale may say.[46]

Such "defensive pessimism" may seem at odds with our model, but in fact it can be captured with a very simple variant. The above are situations where the underlying motive for information-manipulation is still the same, namely to alleviate the shirking incentives of future selves. The only difference is that ability –or more generally the signal received about ability– is a *substitute* rather than a complement to effort in generating future payoffs. This clearly gives the agent

are relevant, i.e. $\beta \in (\underline{\beta}, \bar{\beta})$ in the notation of Proposition 2. The value $R(\beta)$ where the curve steps up is where the two effects just offset each other, and mixed strategies may occur. It shifts with β as indicated on the figure. When repression is costly ($a > 0 = b$) the best response locus starts with a positive value at $r = q$, then (after a possible initial decline), rises towards 1 at $r = R(\beta)$. To the right of this point, it is as on Figure 3.

[45] More generally, if Self 1's beliefs r^* are independent of the actual strategy pursued, the equilibrium is simple and uniquely determined: $\lambda = 0$ for $r^* < R(\beta)$, while $\lambda = 1$ for $r^* > R(\beta)$; the case $r^* = R(\beta)$ has measure zero.

[46] In an interpersonal context such self–deprecation may just be strategic "fishing for compliments" and reassurance–seeking; see Benabou and Tirole (1999). We leave this motive aside in the present paper.

an incentive to discount, ignore and otherwise repress signals of high ability, such as previous achievements, as these would increase the temptation to "coast" or "slack off".

The substitutability may arise directly in the performance "production function" which, instead of the multiplicative form $\gamma(e,\theta) = \theta V e$ which we have assumed, may be of the form $\gamma(e,\theta)$ with $\gamma_{e\theta} < 0$. More interestingly, it will typically occur when the *reward for performance* is of a "pass-fail" nature, such as in obtaining a diploma, making a sale, being hired or fired (tenure, partnership) –perhaps also in marriage and divorce. To formalize this argument, let us maintain the original assumption that performance is multiplicative in ability and effort: $\gamma(\theta,e,\varepsilon) = \varepsilon\theta e$, where ε is some random disturbance (luck) with c.d.f. $H(\varepsilon)$. The payoff V, however, is now conditional on performance exceeding some cutoff level $\bar{\gamma}$. Self 1's utility function is thus

$$\beta\delta V \Pr\left[\varepsilon\theta e \geq \bar{\gamma}\right] - ce = \beta\delta V\left(1 - H(\bar{\gamma}/\theta e)\right) - ce. \tag{31}$$

It can easily be verified that if the density $h = H'$ is such that $xh'(x)/h(x) > -1$ on the relevant range of $x \equiv \bar{\gamma}/\theta e$, the optimal effort is decreasing in θ.

Finally, the substitutability could occur not between e and θ, but between e and the signal received on θ. Suppose for instance that prior to date zero, the individual has achieved a level of performance $v \in \{v_L, v_H\}$, with $v_H > v_L$. Assume further that if the individual does not exert effort at date one, his final (cumulative) payoff will be that same v. If he does exert effort and is successful, he will obtain $V > v_H$. If he does not exert effort, he may still succeed and achieve V with some lower probability; otherwise he will receive his fallback level v. In such a situation, an individual who at date 0 is uncertain as to whether he has achieved v_L or v_H has conflicting incentives as to what he would like to believe at date 1. On one hand, having achieved v_H rather than v_L is likely to be a positive signal about his ability, which tends to improve motivation. On the other hand, it may be that conditional on v_H, success is more of "a sure thing" (less related to effort) than conditional on v_L. If the latter effect dominates, so that

$$\begin{aligned} 0 < \Pr\left[\text{success}|e=\mathbf{1},v_H\right] - \Pr\left[\text{success}|e=0,v_H\right] &< c/\beta\delta V \\ &< \Pr\left[\text{success}|e=1,v_L\right] - \Pr\left[\text{success}|e=0,v_L\right], \end{aligned}$$

Self 0 would prefer Self 1 to believe that v_L occurred rather than v_H, so that he not be tempted to "rest on his laurels".

5.2 Self-esteem as a consumption good

We have until now emphasized the fundamental value of self-confidence as a motivator for effort and other tasks on which the individual may be tempted to shirk. This approach provides an explanation of *why* people care about their self-image, as well as of *how* and *how much* they

care about it: the value of self-confidence arises endogenously from fundamental preferences, technological constraints, and the structure of incentives.

An alternative motive advanced by psychologists for self-esteem maintenance involves purely affective (rather than functional) concerns: people just *like* to think of themselves as good, able, generous, attractive, etc. Formally, self-image is simply posited to be an argument of the utility function. This potentially allows people to care about a broader set of self-attributes than a purely motivation-based theory: they may for instance want to perceive themselves as honest and compassionate individuals, good citizens, faithful spouses ... or, on the contrary, pride themselves on being ruthless businessmen, ultra-rational economists, irresistible seducers, etc. There is somewhat of an embarrassment of riches here, with few constraints on what arguments should enter the utility function, and with what sign.

Let us therefore focus, as before, on the trait of "general ability", which presumably everyone views as a good. This is also the type of attribute from which agents are assumed to derive consumption value in Weinberg (1999) and Köszegi (1999), as well as in some interpretations which Akerlof and Dickens (1982) offered for their model of dissonnance reduction. The tradeoff between acquiring information for decision-making purposes and not acquiring it (or distorting it) in order to protect self-esteem can then be modelled in a static fashion, by positing preferences of the form:

$$\mathsf{E}\left[\max\left\{\theta V - c, 0\right\} + u(\theta)\right], \tag{32}$$

where θ denotes the individual's self-perceived ability at the time of the effort decision. The first term always generates a demand for accurate information. If the valuation of self-esteem $u(\theta)$ is concave, however, there arises a risk-aversion with respect to self-relevant signals. The individual may then, once again, turn down free information or even engage in self-handicapping. In a similar spirit, Weinberg (1999) shows that he may choose easy tasks with a low expected payoff, because these are less revealing about ability.[47] There is of course less evidence on the shape of the "reflective" preference function $u(\theta)$ than on hyperbolic discounting and people's general tendency to procrastinate. It is thus equally likely that the function $u(\theta)$ is convex, at least over some range; for instance, it could be that only significant boosts to self-esteem make

[47] In a different context, Rabin (1995) makes beliefs about the negative externalities of one's actions (on other people, animals, or the environment) an argument of the utility function, and assumes concavity. This provides an explanation for why people may prefer not to know of the potential harm caused by their consumption choices. Such behavior may be interpreted as a form of self–esteem management, where self–esteem is derived from seeing oneself as a "moral" person. Caplin and Leahy (1999) study a general class of preferences where initial perceptions of future lotteries enter into the intertemporal utility function. Depending on whether the dependence is concave or convex, a person will choose to avoid information that would make the future lottery more risky (e.g., precommit to a vacation destination because the risk that he may change his mind reduces the enjoyment derived from anticipation) or, on the contrary, seek out information or situations that increase the stakes (e.g., betting on one's favorite team).

the individual feel measurably better off. In such cases he will be an avid information-seeker, choosing tasks that are excessively hard or risky but very informative, as a way of "gambling for resurrection".

Another one of our main results which carries over to the case of self–esteem as a consumption good is that memory management and similar forms of equilibrium self-deception may be ultimately detrimental (recall the quotation from Plato), while conversely personal rules not to tamper with the encoding and recall of information, such as Darwin's, can be valuable. The basic insight is that of *externalities across information states* (good and bad signals), which arise as long as agents are Bayesian –or at least reasonably aware of their own incentives to distort signals. As explained earlier, having only good news is not such a great boost to self-esteem once the agent realizes that he had reasons to bias his information-gathering or information-storing strategy. To see that the welfare implications under (32) are broadly similar to those derived in Section 3.5, consider for instance an agent who, when faced with bad news, is just indifferent between remembering them and repressing them: the decision-making and self-esteem motives in (32) just offset each other, leading to a mixed strategy. Relative to a truthful strategy ($\lambda \equiv 1$), he achieves zero expected gains in states of the world where $\sigma = L$. When $\sigma = \emptyset$ he does strictly less well, both because his self-esteem $u(\theta)$ is not as high due to the self-distrust motive, and because his effort decision must rely on more noisy information. On average, his ex-ante welfare is thus lower than if he could have bound himself to always face the truth, or followed that strategy as part of an alternative equilibrium (when it exists).

6 Concluding Comments

Building on several themes developed in social psychology, we proposed in this paper a general economic model of why and how people care about their own self-image. We then showed how this concern for the maintenance of self-esteem may lead them to engage in seemingly irrational behavior such as avoiding information, creating obstacles to their own performance, and trying to "deceive themselves" through elaborate strategies of selective memory, limited awareness, and related forms of belief manipulation.

We discussed several examples of economic applications along the way, but many more quickly come to mind when thinking of the role played by self-perceived ability in a person's educational, financial, or labor market decisions. Others are best analyzed in the context of *inter*personal interactions, such as principal-agent or bargaining relationships. In such strategic situations, conflicting interests will generally imply that self-confidence matters even when agents are fully time-consistent. These issues are explored in our companion paper (Benabou and Tirole 1999).

APPENDIX

Proof of Proposition 2 For all r and β in $[0,1]$, let us define:

$$\chi(r,\beta) \equiv \int_{\beta\delta\theta_L}^{\beta\delta\theta(r)} (\delta\theta_L - c)\, d\Phi(c), \tag{A.1}$$

which, up to a factor of $\beta\delta$, measures the incentive to forget bad news, $U_C(\theta_L|\, r^*) - U_T(\theta_L)$.

Lemma 1 *For all $r \in [0,1]$, there exists a unique $B(r) \in [0,1]$ such that $\chi(r, B(r)) = 0$ and:*
i) $\chi(r,\beta) > 0$ for all $\beta < B(r)$, while $\chi(r,\beta) < 0$ for all $\beta > B(r)$;
ii) $B(r) > \theta_L/\theta(r)$, and $B(r)$ is strictly decreasing in r.

Proof. For any given r, it is clear from (A.1) that $\chi(r,\beta) > 0$ for $\beta \in [0, \theta_L/\theta(r)]$, while $\chi(r,1) < 0$. Moreover, for all $\beta > \theta_L/\theta(r)$, we have:

$$\frac{\partial\chi(r,\beta)}{\partial\beta} = \delta^2\theta(r)\,[\theta_L - \beta\theta(r)]\,\varphi(\beta\delta\theta(r)) - \delta^2\theta_L\,[\theta_L - \beta\theta_L]\,\varphi(\beta\delta\theta_L) < 0. \tag{A.2}$$

This establishes the existence and uniqueness of the root $B(r) \in [\theta_L/\theta(r), 1]$. Moreover,

$$\frac{\partial\chi(r,\beta)}{\partial r} = \beta\delta^2\,(\theta_H - \theta_L)\,[\theta_L - \beta\theta(r)]\,\varphi(\beta\delta\theta(r)), \tag{A.3}$$

so $\partial\chi(r, B(r))/\partial r < 0$ since $B(r)\theta(r) > \theta_L$. Therefore, by the implicit function theorem, $B'(r) < 0$ for all r. ||

To conclude the proof of Proposition 2, consider the following cases.

a) For $\beta \geq B(q)$ we have, for all $r \in [q,1]$, $\beta > B(r)$ and therefore $\chi(r,\beta) < 0$. Memorizing bad news is thus the optimal strategy, which establishes claim (i) of the Proposition.

b) For $\beta \leq B(1)$ we have, for all $r \in [q,1]$, $\beta < B(r)$ and therefore $\chi(r,\beta) > 0$. Forgetting bad news is thus the optimal strategy, which establishes claim (iii).

c) For $\beta \in (B(1), B(q))$ there exists by the lemma a unique inverse function $R(\beta) \equiv B^{-1}(\beta)$, such that $\chi(R(\beta),\beta) = 0$. Moreover, the function R is decreasing, and for any $r \in (q,1)$, $\chi(r,\beta)$ has the sign of $R(\beta) - r$. This implies that the only equilibrium with $r < R(\beta)$ is $r = q$ ($\lambda = 0$), with $\chi(q,\beta) > 0$; the only equilibrium with $r > R(\beta)$ is $r = 1$, ($\lambda = 1$), with $\chi(1,\beta) < 0$; $r = R(\beta)$ is an equilibrium, which corresponds to $\Lambda(\beta) = (1 - q/R(\beta))\,(1-q)$. Defining $\underline{\beta} \equiv B(1)$ and $\bar{\beta} \equiv B(q)$ concludes the proof of the proposition. ■

Proof of Proposition 3 We shall solve for equilibria in terms of the reliability of memory, r^*, as this is analytically simpler; the recall strategy λ^* is then obtained by inverting (15).

Let us first rewrite the incentive to forget, given by (26), as:

$$\psi(r,\beta) = r\left(\Delta\theta\right)\left(\frac{\beta^2\delta^3}{\bar{c}}\right)\left((1-\beta)\theta_L - \frac{\beta r}{2}\left(\Delta\theta\right)\right) r - ar\left(\frac{1-q}{r-q}\right) + br\left(\frac{1-q}{q(1-r)}\right).$$

Defining, for all $\beta \in [0,1]$:

$$R(\beta) \equiv \left(\frac{1-\beta}{\beta}\right)\frac{2\theta_L}{\Delta\theta}, \tag{A.4}$$

$$(\beta) \equiv (\Delta\theta)^2\left(\frac{\beta^3\delta^3}{2\bar{c}}\right)\left(\frac{q}{1-q}\right). \tag{A.5}$$

it is clear that $\psi(r,\beta) \geq 0$ if and only if:

$$P(r,\beta) \equiv (\beta)\left(R(\beta)-r\right)(r-q) \geq aq - b\left(\frac{r-q}{1-r}\right). \tag{A.6}$$

Multiplying by $(1-r)$ shows that the sign of $\psi(r,\beta)$ is given by that of a third-degree polynomial in r. Let us now specialize (A.6) to the case where remembering is costless but forgetting or repressing costly, $b=0$. Solving $\psi(r,\beta)=0$ then reduces to looking for the intersections of the *quadratic* polynomial $P(r,\beta)$ with the horizontal line aq. There are several cases to consider.

1) For $R(\beta) < q$, or equivalently $\beta > R^{-1}(q) \equiv \beta_3$, it is clear that $P(r,\beta) < 0$ on $[q,1]$, therefore the only equilibrium is $r=1$.

2) For $q < R(\beta) \equiv \beta_3$ the polynomial $P(r,\beta)$ is positive on $r \in [q, R(\beta)]$, and negative outside. Let $\bar{a} \equiv q^{-1}\max_{r\in[q,1]}\left(P(r,\beta)\right) > 0$, and $\underline{a} \equiv P(1,\beta) \leq \bar{a}$.

a) If $a > \bar{a}$, then $P(r,\beta) < 0$ on $[q, R(\beta)]$, so the only equilibrium is again $r=1$.

b) If $a \leq \bar{a}$, the equation $P(r,\beta) = aq$ has two roots $r_1(a)$ and $r_2(a)$, both in the interval $[q, R(\beta)]$, with $r_1(a) \leq r_2(a)$, r_1 decreasing and r_2 increasing. To these one can associate two functions, $\lambda_1(a)$ and $\lambda_2(a)$, by inverting (15). Let us now distinguish the following subcases.

i) For $q < R(\beta) < 1$, or equivalently $\beta_2 \equiv R^{-1}(1) < \beta < R^{-1}(q) = \beta_3$, both $r_1(a)$ and $r_2(a)$ are in $(q, R(\beta))$ and represent equilibria. On $[q, r_1(a))$ and $(r_2(a), 1]$ we have $P(r,\beta) < aq$, hence $\psi(r,\beta) < 0$. This means that the only other equilibrium is $r=1$.

ii) For $1 < R(\beta) < 2-q$, or equivalently $\beta_1 \equiv R^{-1}(2-q) < \beta < \beta_2 = R^{-1}(1)$, the polynomial $P(r,\beta)$ reaches its maximum at $(q + R(\beta))/2 < 1$. Thus $P(r,\beta)$ is positive and hill-shaped on $[q,1]$, and $\underline{a} \equiv P(1,\beta) > 0$. This implies that for $\underline{a} < a < \bar{a}$ we have $q < r_1(a) < r_2(a) < 1$, while for $a < \underline{a}$ we have $q < r_1(a) < 1 < r_2(a)$. In the first case the equilibria are $r \in \{r_1(a), r_2(a), 1\}$, as in case (i) above. In the latter situation the only equilibrium is $r = r_1(a)$.

iii) For $2-q < R(\beta)$, or equivalently $\beta < \beta_1 \equiv R^{-1}(2-q)$, the polynomial $P(r,\beta)$ is strictly increasing on $[q,1]$, so the only equilibrium is $r = r_1(a)$ whenever $a < \underline{a} \equiv P(1,\beta)$. It is $r=1$ whenever $a \geq \underline{a}$. ■

Proof of Proposition 4 When bad news are systematically forgotten, $\lambda^* = 1$, and therefore $r^* = q$. Computing $\Delta W(0,q)$ from (29) with $M=1$ yields:

$$\begin{aligned}
\Delta W(0,q) &= (1-q)\int_{\beta\delta\theta_L}^{\beta\delta\theta(q)}(\delta\theta_L - c)d\Phi(c) - q\int_{\beta\delta\theta(q)}^{\beta\delta\theta_H}(\delta\theta_H - c)d\Phi(c) \\
&= q\int_0^{\beta\delta\theta(q)}(\delta\theta_H - c)d\Phi(c) + (1-q)\int_0^{\beta\delta\theta(q)}(\delta\theta_L - c)d\Phi(c) \\
&\quad -q\int_0^{\beta\delta\theta_H}(\delta\theta_H - c)d\Phi(c) - (1-q)\int_0^{\beta\delta\theta_L}(\delta\theta_L - c)d\Phi(c) \\
&= \int_0^{\beta\delta\theta(q)}\left[\delta(q\theta_H + (1-q)\theta_L) - c\right]d\Phi(c) \\
&\quad -q\int_0^{\beta\delta\theta_H}(\delta\theta_H - c)d\Phi(c) - (1-q)\int_0^{\beta\delta\theta_L}(\delta\theta_L - c)d\Phi(c),
\end{aligned}$$

or, finally:

$$\Delta W(0,q) = \beta^{-1}\left[\Gamma(\beta\delta\left(q\theta_H + (1-q)\theta_L\right),\beta) - q\Gamma(\beta\delta\theta_H,\beta) - (1-q)\Gamma(\beta\delta\theta_L,\beta)\right], \tag{A.7}$$

hence the first result. Note that $\beta\delta\Delta W(0,q)$ is (minus) the ex-ante value of information, i.e. of always knowing the true θ rather than have only the uninformed prior or posterior $\theta(q)$. The second result follows from the fact that $\partial^2\Gamma(Z,\beta)/\partial Z^2 = (2-\beta)\varphi(Z) + (1-\beta)\,Z\varphi'(Z)$. ■

Welfare Rankings of Multiple Equilibria. As stated following Proposition 4, we construct here a simple example where both $\lambda^* = 0$ and $\lambda^* = 1$ are equilibria (simultaneously), and where *either one* –depending on parameter values– may lead to higher ex-ante welfare.

First, let $\theta_L < \theta_H$ and $q \in (0,1)$, so that $\theta_L < \theta(q) = q\theta_H + (1-q)\theta_L < \theta_H$. For $\beta < 1$ but not too small we have $\beta\theta_L < \theta_L < \beta\theta(q) < \theta(q) < \beta\theta_H < \theta_H$. Next, let the date–1 cost take two values: $c \in \{\underline{c}, \bar{c}\}$, with $\underline{c}/\delta \in (\beta\theta_L, \theta_L)$, $\bar{c}/\delta \in (\theta(q), \beta\theta_H)$ and $\pi \equiv \Pr[c = \underline{c}] \in (0,1)$. The $\lambda^* = 0$ strategy is then always an equilibrium, since $\psi(q,\beta)/\beta\delta = \pi(\delta\theta_L - \underline{c}) > 0$. As to $\lambda^* = 1$, it is also an equilibrium whenever $\psi(1,\beta)/\beta\delta = \pi(\delta\theta_L - \underline{c}) - (1-\pi)(\bar{c} - \delta\theta_L) < 0$, or

$$\frac{\pi}{1-\pi} < \frac{\bar{c} - \delta\theta_L}{\delta\theta_L - \underline{c}} \equiv \rho, \tag{A.8}$$

which we shall assume. For all such values of π there are thus two pure strategy equilibria (with a mixed-strategy one in between, which is always inferior to the truthful-recall equilibrium since

$M \equiv 0$). The memory-manipulation equilibrium $\lambda^* = 0$ yields higher ex-ante welfare when

$$\Delta W(0,\beta) = (1-q)\pi(\delta\theta_L - \underline{c}) - q(1-\pi)(\delta\theta_H - \bar{c}) > 0,$$

or:

$$\frac{\pi}{1-\pi} > \left(\frac{q}{1-q}\right)\left(\frac{\delta\theta_H - \bar{c}}{\delta\theta_L - \underline{c}}\right) \equiv \rho'. \tag{A.9}$$

Since $\theta(q) < \delta\bar{c}$, it is easily verified that $\rho' < \rho$. Thus for $\pi/(1-\pi) \in (\rho', \rho)$, the $\lambda = 0$ equilibrium is ex-ante superior to the one with $\lambda = 1$. For $\pi/(1-\pi) < \rho'$ the reverse is true. ■

References

[1] Akerlof, G. and Dickens, W. (1982) "The Economic Consequences of Cognitive Dissonance," *American Economic Review*, 72(3): 307–319.

[2] Ainslie, G. (1992) *Picoeconomics : The Strategic Interaction of Successive Motivational States Within the Person (Studies in Rationality and Social Change)*. Cambridge, England and New York: Cambridge University Press.

[3] Ainslie, G. (1999) "Breakdown of Will," mimeo, May.

[4] Alloy, L.B. and Abrahamson, L.Y. (1979) "Judgement of Contingency in Depressed and Nondepressed Students: Sadder but Wiser?" *Journal of Experimental Psychology: General*, 108, 441–485.

[5] Arkin, R.M. and A. H. Baumgardner (1985) "Self-Handicapping," in *Attribution: Basic Issues and Applications*, edited by J. Harvey and G. Weary, New York: Academic Press.

[6] —(1986) "Self-Presentation and Self-Evaluation: Processes of Self-Control and Social Control ," in *Public Self and Private Self*, edited by R. Baumeister, New York: Springer Verlag.

[7] Bandura, A. (1977) *Self Efficacy: The Exercise of Control*. W. H. Freeman Company.

[8] Baumeister, R. (1998) "The Self," in *The Handbook of Social Psychology*, edited by D. Gilbert, S. Fiske and G. Lindzey, Boston: McGraw-Hill.

[9] Benabou, R. and J. Tirole (1999) "Self–Confidence and Social Interactions." IDEI mimeo, June.

[10] Berglas, S. and E. Jones (1978) "Drug Choice as a Self-handicapping Strategy in Response to Non-Contingent Success," *Journal of Personality and Social Psychology*, 36: 405–417.

[11] Brocas, I. and Carrillo, J. (1999) "Entry Mistakes, Entrepreneurial Boldness and Optimism," ULB–ECARE mimeo, June.

[12] Caplin, A. and Leahy, J. (1999) "Psychological Expected Utility Theory," New York University mimeo, May.

[13] Carrillo, J., and T. Mariotti (1997) "Strategic Ignorance as a Self-Disciplining Device," forthcoming in *Review of Economic Studies*.

[14] Crary, W.G. (1966) "Reactions to Incongruent Self–Experiments," *Journal of Consulting Psychology*, 30, 246–252.

[15] Darley, J. and G. Goethals (1980) "People's Analyzes of the Causes of Ability-Linked Performances," in *Advances in Experimental Social Psychology*, vol. 13, ed. L. Berkowicz, New York: Academic Press.

[16] Deci, E. (1975) *Intrinsic Motivation*, New York: Plenum.

[17] Dewatripont, M., Jewitt, I., and J. Tirole (1999a)"The Economics of Career Concerns, Part I: Comparing Information Structures," *Review of Economic Studies*, 66(1): 183–198.

[18] —(1999b) "The Economics of Career Concerns, Part II: Application to Missions and Accountability of Government Agencies," *Review of Economic Studies* , 66(1): 199–217.

[19] Elster, J. (1999) "Motivated Belief Formation," Columbia University mimeo, June.

[20] Fazio, R., and M. Zanna (1981) "Direct Evidence and Attitude-Behavior Consistency," in L. Berkowitz (ed.) *Advances in Experimental Social Psychology*, vol. 14, New York: Academic Press.

[21] Festinger, L. (1954) "A Theory of Social Comparison Processes," *Human Relations*, 7: 117–140.

[22] Fingarette, H. (1985) "Alcoholism and Self-Deception," in *Self-Deception and Self-Understanding*, ed. by M. Martin, University Press of Kansas.

[23] Freud, S. (1927) *The Ego and the Id*, London: Hogarth.

[24] —(1938) *A General Introduction to Psychoanalysis*. New York: Garden City Publishing Co.

[25] Frey, D. (1981) "The Effect of Negative Feedback about Oneself and Cost of Information on Preference for Information about the Source of this Feedback," *Journal of Experimental Social Psychology*, 17: 42–50.

[26] Fudenberg, D. and J. Tirole (1991) *Game Theory*. Cambridge: MIT Press.

[27] Gilbert, D. and J. Cooper (1985) "Social Psychological Strategies of Self- Deception," in M. Martin, ed. *Self-Deception and Self-Understanding*, University Press of Kansas.

[28] Gilbert D. and D. Silvera (1996) "Overhelping," *Journal of Personality and Social Psychology*, 70: 678-690.

[29] Gilovich, T. (1991) *How We Know What Isn't So*. New York: Free Press.

[30] Greenier, K., Kernis, M. and Wasschull, S. (1995) "Not All High (or Low) Self–Esteem People Are the Same: Theory and Research on the Stability of Self–Esteem", in *Efficacy, Agency and Self–Esteem,* M. Kernis ed., New York: Plenum Press

[31] Greenwald, A. (1980) "The Totalitarian Ego: Fabrication and Revision of Personal History," *American Psychology*, 35: 603–613.

[32] Gur, R. and H. Sackeim (1979) "Self-Deception: A Concept in Search of a Phenomenon," *Journal of Personality and Social Psychology*, 37: 147–169.

[33] Heider, F. (1958) *The Psychology of Interpersonal Relations*. New York: Wiley.

[34] Holmström,B. (1999) "Managerial Incentive Problems: A Dynamic Perspective," *Review of Economic Studies*, 66(1): 169–182.

[35] Hull, J.G. and R. D. Young, (1983) "The Self-Awareness-Reducing Effects of Alcohol: Evidence and Implications," in *Psychological Perspectives on the Self*, vol. 2, edited by J. Suls and A. Greenwald, Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

[36] James, W. (1890) *The Principles of Psychology*. Cleveland, OH: World publishing.

[37] Jones, E., Rhodewalt, F., Berglas, S. and J. Skelton (1981) "Effects of Strategic Self-Presentation on Subsequent Self-Esteem," *Journal of Personality and Social Psychology*, 41: 407–421.

[38] Kelley, H. (1972) in E. Jones et al, eds. *Attribution: Perceiving the Causes of Behavior* Morristown, NJ: General Learning Press.

[39] Kolditz, T. and R. Arkin (1982) "An Impression Management Interpretation of the Self-Handicapping Strategy," *Journal of Personality and Social Psychology*, 43: 492–502.

[40] Korner, I. (1950) "Experimental Investigation of Some Aspects of the Problem of Repression: Repressive Forgetting." New York, NY: Contributions to Education, No. 970, Bureau of Publications, Teachers' College, Columbia University.

[41] Köszegi, B. (1999) "Self–Image and Economic Behavior," MIT mimeo, October.

[42] Kuiper, N.A. and Derry, P.A. (1982) "Depressed and Nondepressed Content Self–Reference in Mild Depression," *Journal of Personality*, 50, 69–79.

[43] Kunda, Z. and Sanitioso, R. (1989) "Motivated Changes in the Self–Concept," *Journal of Personality and Social Psychology*, 61, 884–897.

[44] Laibson, D. (1997) "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, 112: 443–478.

[45] Leary, M. and Downs, D. (1995) "Interpersonal Functions of the Self–Esteem Motive: The Self–Esteem System as Sociometer", in *Efficacy, Agency and Self–Esteem*, M. Kernis ed., New York: Plenum Press.

[46] Laughlin, H.P. (1979) *The Ego and Its Defenses*, The National Psychiatric Endowment Fund eds., second edition. New York, NY: HJason Aaronson, Inc.

[47] Loewenstein, G. and Prelec, D. (1992) "Anomalies in Intertemporal Choice: Evidence and Interpretation," *Quarterly Journal of Economics*, 107(2): 573–597.

[48] Mele, A. "Real Self–Deception," *Behavioral and Brain Sciences*, 20(1): 91–136.

[49] Miller, R., Brickman, P. and D. Bolen (1975) "Attribution vs Persuasion as a Means for Modifying Behavior," *Journal of Personality and Social Psychology*, 31: 430–441.

[50] Mischel, W., Ebbesen, E.B. and Zeiss, A.R. (1976) "Determinants of Selective Memory about the Self," *Journal of Consulting and Clinical Psychology*, 44, 92–103.

[51] Mullainathan, S. (1998) "A Memory Based Model of Bounded Rationality," mimeo, MIT.

[52] Mulligan, C. (1996) "A Logical Economist's Argument Against Hyperbolic Discounting," University of Chicago mimeo, February.

[53] Murray, S.L. and Holmes, J.G. (1994) "Seeing Virtues in Faults: Negativity and the Transformation of Interpersonal Narratives in Close Relationships," *Journal of Personality and Social Psychology*, 20, 650–663.

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

[55] Nisbett, R. and T. Wilson (1977) "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review*, 84: 231–259.

[56] O'Donoghue, T. and Rabin, M. (1999) "Doing it Now or Later," *American Economic Review, 89(1), 103–124.*

[57] Osborne, M. and A. Rubinstein (1994) *A Course in Game Theory*. Cambridge: MIT Press.

[58] Phelps, E. and Pollack, R. (1968) "On Second–Best National Savings and Game–Equilibrium Growth," *Review of Economic Studies, 35: 185–199.*

[59] Rabin, M. (1995) "Moral Preferences, Moral Rules, and Belief Manipulation," University of California mimeo, April.

[60] Rabin, M. and Schrag, (1999) "First Impressions Matter: A Model of Confirmatory Bias," *Quarterly Journal of Economics, 114, 37–82.*

[61] Rhodewalt, F.T. (1986) "Self-Presentation and the Phenomenal Self: On the Stability and Malleability of Self-Conceptions," in *Public Self and Private Self*, edited by R. Baumeister, New York: Springer Verlag.

[62] Salancik, G. (1977) "Commitment and the Control of Organizational Behavior and Belief," in *New Directions in Organizational Behavior*, edited by B. Staw and G. Salancik, Chicago: St. Clair Press.

[63] Schacter, D. (1996) *Searching for Memory*. Basic Books.

[64] Schlenker, B. (1986) "Self-identification: Toward an Integration of the Private and Public Self," in *Public Self and Private Self*, ed. by R. Baumeister, New York: Springer Verlag.

[65] Sartre (1953) *The Existential Psychoanalysis*, (H.E. Barnes, trans.). New York: Philosophical Library.

[66] Seligman, E. (1990) *Learned Optimism: How to Change Your Mind and Your Life*. New York: Simon and Schuster.

[67] Smith, E. (1998) "Mental Representation and Memory," in *Handbook of Social Psychology*, edited by D. Gilbert, S. Fiske, and G. Lindzey, Boston: McGraw Hill.

[68] Snyder, C., Higgins, R., and R. Stucky (1983) *Excuses: Masquerades in Search of Grace*. New York: John Wiley.

[69] Snyder, C. (1985) "Collaborative Companions: The Relationship of Self-Deception and Excuse Making," in M. Martin, ed. *Self-Deception and Self-Understanding*, University Press of Kansas.

[70] Strotz, R. (1956) "Myopia and Inconsistency in Dynamic Utility Maximization," *Review of Economic Studies*, 23: 165–180.

[71] Swann, W.B. Jr. (1983) "Self-Verification: Bringing Social Reality into Harmony with the Self," in *Psychological Perspectives on the Self*, vol. 2, edited by J. Suls and A. Greenwald, Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers, pp33–66.

[72] Swann, W.B. Jr. (1996) *Self Traps: the Elusive Quest for Higher Self–Esteem.* New York: W.H. Freeman and Company.

[73] Taylor, S.E. and Brown, J.D. (1988) "Illusion and Well–Being: A Social Psychological Perspective on Mental Health," *Psychological Bulletin*, 103–193–210.

[74] Tedeschi, J.T. (1986) "Private and Public Experiences and the Self," in *Public Self and Private Self*, ed. by R. Baumeister, New York: Springer Verlag.

[75] Tesser, A. and J. Moore (1986) "On the Convergence of Public and Private Aspects of Self," in *Public Self and Private Self*, ed. by R. Baumeister, New York: Springer Verlag.

[76] Weinberg, B. "A Model of Overconfidence," Ohio State University mimeo, August.

[77] Weinstein, N. (1980) "Unrealistic Optimism About Future Life Events," *Journal of Personality and Psychology*, 39(5): 806–820.

[78] Wrangham, R. (1999) "Is Military Incompetence Adaptive?", mimeo, Harvard University.

[79] Zuckerman, M. (1979) "Attribution of Success and Failure Revisited, or the Motivational Bias is Alive and Well in Attribution Theory," *Journal of Personality*, 47: 245–87.

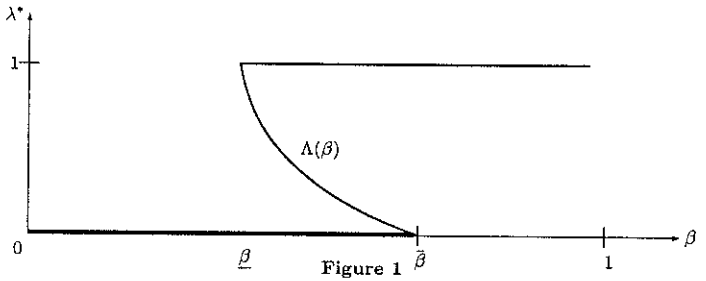

Figure 1

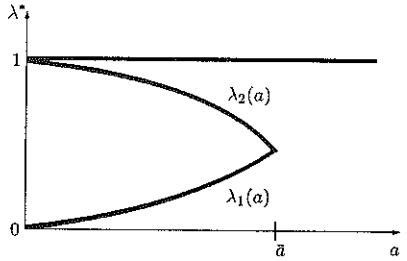

Figure 2.1: Case $\beta_2 < \beta < \beta_3$

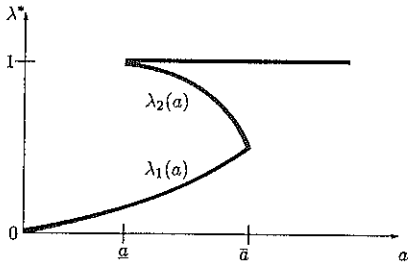

Figure 2.2: Case $\beta_1 < \beta < \beta_2$

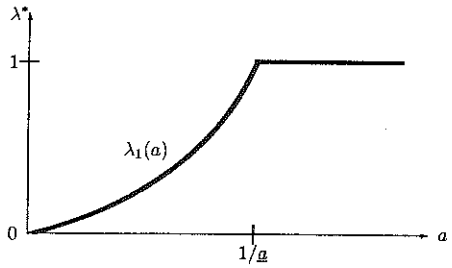

Figure 2.3: Case $\beta < \beta_1$

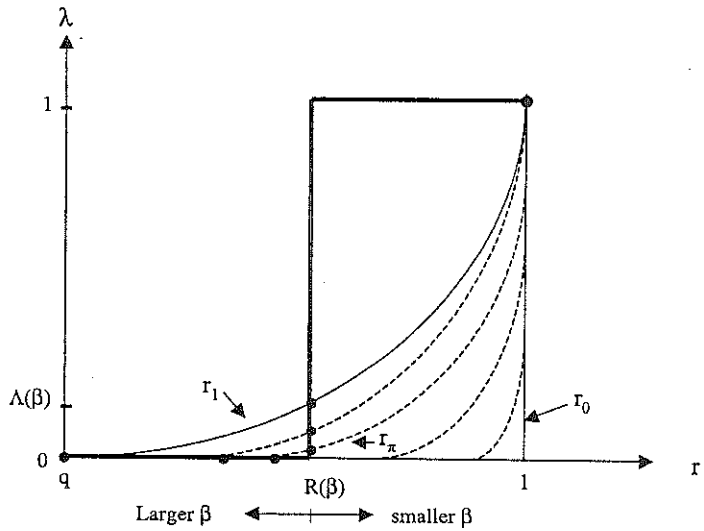

Figure 3: Equilibrium set for varying degrees of Bayesian rationality.