SELF-CONFIDENCE AND PERSONAL MOTIVATION*

ROLAND BÉNABOU AND JEAN TIROLE

We analyze the value placed by rational agents on self-confidence, and the strategies employed in its pursuit. Confidence in one's abilities generally enhances motivation, making it a valuable asset for individuals with imperfect willpower. This demand for self-serving beliefs (which can also arise from hedonic or signaling motives) must be weighed against the risks of overconfidence. On the supply side, we develop a model of self-deception through endogenous memory that reconciles the motivated and rational features of human cognition. The resulting intrapersonal game of strategic communication typically leads to multiple equilibria. While "positive thinking" can improve welfare, it can also be self-defeating (and nonetheless pursued).

> 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, *Beyond Good and Evil*].
>
> 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 in *The Life of Charles Darwin*, by Francis Darwin].

INTRODUCTION

The maintenance and enhancement of self-esteem has always been identified as a fundamental human impulse. Philosophers, writers, educators, and of course psychologists all have emphasized the crucial role played by self-image in motivation, affect, and social interactions. The aim of this paper is to bring

* This paper was previously titled "Self-Confidence: Intrapersonal Strategies [1999]." For helpful comments and discussion we are grateful to Dilip Abreu, Olivier Blanchard, Isabelle Brocas, Edward Glaeser, Daniel Gilbert, Ian Jewitt, David Laibson, George Loewenstein, Andrew Postlewaite, Marek Pycia, Matthew Rabin, Julio Rotemberg, and three anonymous referees. We also thank conference and seminar participants at the University of Chicago, Columbia University, Cornell University, the Massachusetts Institute of Technology, the National Bureau of Economic Research, Northwestern University, New York University, the Oxford Young Economists' Conference at Oxford University, the University of Pennsylvania, Princeton University, Stanford University, and Yale University. Bénabou gratefully acknowledges financial support from the National Science Foundation (SES-0096431).

The Quarterly Journal of Economics, August 2002

these concerns into the realm of economic analysis, and show that this has important implications for how agents process information and make decisions. Conversely, the tools of economic modeling can help shed light on a number of apparently irrational behaviors documented by psychologists.

Indeed, both the demand and the supply sides of self-confidence appear at odds with economists' view of human behavior and cognition. Why should people prefer rosy views of themselves to accurate ones, or want to impart such beliefs to their children? From car accidents, failed dot.com firms, and day trading to the space shuttle disaster and lost wars, the costs of overconfidence are plain for all to see. Even granting that some "positive illusions" could be desirable, is it even possible for a rational, Bayesian individual to deceive himself into holding them? Finally, the welfare consequences of so-called self-serving beliefs are far from clear: while "thinking positive" is often viewed as a good thing, self-deception is not, even though the former is only a particular form of the latter.

To analyze these issues, we develop a simple formal framework that unifies a number of themes from the psychology literature, and brings to light some of their economic implications. We first consider the demand side of self-confidence, and identify in Section I three main reasons why people may prefer optimistic self-views to accurate ones: a consumption value, a signaling value, and a motivation value. First, people may just derive utility from thinking well of themselves, and conversely find a poor self-image painful. Second, believing—rightly or wrongly—that one possesses certain qualities may make it easier to convince others of it. Finally, confidence in his abilities and efficacy can help the individual undertake more ambitious goals and persist in the face of adversity. While we shall mostly focus on this last explanation, all three should be seen as complementary, and for many purposes work equally well with the supply side of our model (self-deception).

The main reason why we emphasize the motivation theory is its substantially broader explanatory power. Indeed, it yields an endogenous value of self-confidence that responds to the situations and incentives which the individual faces, in a way that can account for both "can-do" optimism and "defensive" pessimism. It also readily extends to economic and social interactions (altruistic or not), explaining why people generally prefer self-confident

partners to self-doubting ones, and invest both time and effort in supporting the latter's morale.

The first premise of the motivation theory is that people have imperfect knowledge of their own abilities, or more generally of the eventual costs and payoffs of their actions.[1] The second one is that ability and effort interact in determining performance; in most instances they are complements, so that *a higher self-confidence enhances the motivation to act.* As demonstrated by the opening quote from James [1890], this complementarity has long been familiar in psychology.[2] It is also consistent with the standard observation that morale plays a key role in difficult endeavors; conversely, when people expect to fail, they fail quite effectively, and failure leads to failure more readily for individuals characterized with low self-esteem [Salancik 1977].

The fact that higher self-confidence enhances the individual's motivation gives anyone with a vested interest in his performance an incentive to build up and maintain his self-esteem. First, the manipulator could be another person (parent, teacher, spouse, friend, colleague, manager) who is eager to see him "get his act together," or otherwise apply himself to the task at hand. Such *interpersonal* strategies are studied in Bénabou and Tirole [2001]. Second, for an individual suffering from time inconsistency (e.g., hyperbolic discounting), the current self has a vested interest in the self-confidence of future selves, as it helps counter their natural tendency to quit too easily. It is in this context, which builds on Carrillo and Mariotti [2000], that we shall investigate a variety of *intrapersonal* strategies of self-esteem maintenance. We shall thus see how and when people may choose to remain ignorant about their own abilities, and why they sometimes deliberately impair their own performance or choose overambitious tasks in which they are sure to fail (self-handicapping).

Section II thus turns to the supply side of the self-confidence

1. The psychology literature generally views introspection as quite inaccurate [Nisbett and Wilson 1977], and stresses that learning about oneself is an ongoing process. Furthermore, the self is constantly changing (e.g., Rhodenwalt [1986]): personal characteristics evolve with age, the goals pursued shift over one's career and life cycle (often as the result of interactions with others), and the personal or economic environment in which these objectives are rewarded is typically variable.

2. Thus, Gilbert and Cooper [1985] note that "the classic attributional model of the causes of behavior . . . [is described by] 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." Additional references are given in Section I. Note, however, that there are also instances where ability and effort are substitutes. As discussed below, we shall consider this case as well.

problem, and the "reality constraints" that limit the extent to which people can engage in wishful thinking. In our model we maintain the standard assumption of individuals as rational (Bayesian) information processors. While almost universal in economics, this view is more controversial in psychology. On one hand, a lot of the classical literature has emphasized rationality and information-seeking in the process of self-identification, documenting the ways in which people update their beliefs according to broad Bayesian principles.[3] On the other hand, the more recent cognitive literature abundantly documents the less rational (or at least, subjectively motivated) side of human inference.

For instance, a substantial body of evidence suggests that people tend to recall their successes more than their failures, and have self-servingly biased recollections and interpretations of their past performances.[4] Similarly, they tend to overestimate their abilities and other desirable traits, as well as the extent to which they have control over outcomes. They also rate their own probabilities as above average for favorable future life events, and below average for unfavorable ones; the more controllable these events through their future actions, the more so.[5]

We shall capture this class of *self-deception* phenomena with a simple game-theoretic model of endogenous memory, or awareness-management, which represents one of the main contributions of this paper. Drawing on evidence about the mechanics and limitations of memory, it shows how to reconcile the motivated ("hot") and rational ("cold") features of human cognition, and could be used in any setting where a demand for motivated beliefs arises. The basic idea is that the individual can, within limits and

3. Thus, *attribution theory* [Heider 1958] emphasizes the distinction between temporary (situational) and enduring (dispositional) characteristics. In economics parlance, the individual filters out noise in order to extract information from past events. In the *social comparison process* [Festinger 1954] individuals assess their ability by comparing their performance with that of people facing similar conditions (familial, cultural, educational, etc.). In other words, 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, and conversely some comfort is derived when others experience adversity (*Schadenfreude*). Relatively sophisticated updating also applies to the interpretation of praise and criticism: a person takes into account not only what others say (or do), but also their possible intentions.

4. 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.

5. See, e.g., Taylor and Brown [1988], Weinstein [1980], Alloy and Abrahamson [1979], and the many other references given in Section II. For recent overviews of the general phenomenon of self-deception, see Gilbert and Cooper [1985] and especially Baumeister [1998] on the psychological evidence, and Elster [1999] and Mele [1999] for the philosophical debates and implications.

possibly at a cost, *affect the probability of remembering* a given piece of data. At the same time, we maintain rational inference, so people realize (at least to some extent) that they have a selective memory or attention.

The resulting structure is that of a game of strategic communication between the individual's temporal selves. In deciding whether to try to repress bad news, the individual weighs the benefits from preserving his effort motivation against the risk of becoming overconfident. Later on, however, he appropriately discounts the reliability of rosy recollections and rationalizations. The implications of this game of asymmetric information are quite different from those of ex ante decisions about information acquisition (e.g., self-handicapping or selective search). In particular, multiple intrapersonal equilibria ("self-traps") may arise, ranging from systematic denial to complete self-honesty. More generally, we characterize the set of perfect Bayesian equilibria and its dependence on the individual's degree of time inconsistency and repression costs ("demand and supply" parameters).

The model also has interesting implications for the distribution of optimism and pessimism across agents, which we examine in Section III. We show that when the costs of repression are low enough, most people typically believe themselves to be more able than they actually are, as well as more able than both the average and the median of the population. A minority will have either realistically low assessments, or actually severely underestimate themselves. We also highlight the key role played by Bayesian-like introspection (understanding, at least partially, one's own incentives for self-esteem maintenance) in the model's results, and why incorporating this essential human trait is more fruitful than modeling agents as naively taking all recollections and self-justifications at face value.

Section IV examines the welfare impact of equilibrium self-deception. Is a more active self-esteem maintenance strategy, when chosen, always beneficial? How can people be "in denial" if it does not serve their best interests? We show that, in addition to the trade-off mentioned earlier between the confidence-maintenance motive and the risks of overconfidence, ex ante welfare reflects a third effect, namely the spoiling of good news by self-doubt. Intuitively, when adverse signals about his ability are systematically repressed, the individual can never be sure that only positive ones were received, even when this is actually true. We characterize the conditions under which always "looking at

the bright side" pays off on average or, conversely, when it would be better to always "be honest with yourself," as Charles Darwin apparently concluded.

In Section V we turn to the case where ability and effort are substitutes rather than complements. This typically occurs when the payoff for success is of a "pass-fail" nature, or characterized by some other form of satiation. Since a high perceived ability may now increase the temptation to exert low effort ("coasting"), this case allows us to account for what psychologists refer to as "defensive pessimism:" the fact that people sometimes minimize, rather than aggrandize, their previous accomplishments and expectations of future success. Another variant of the model considered in this section involves replacing the motivation value of self-confidence with a purely affective one. Section VI concludes the paper. All proofs are gathered in the Appendix.

This paper is related to several strands of the new literature that tries to better link economics and psychology. A hedonic concern for self-image, in the form of preferences over beliefs, was first explored in Akerlof and Dickens' [1982] well-known model of dissonance reduction, and more recently in Rabin [1995], Weinberg [1999], and Köszegi [1999]. In emphasizing an endogenous value of self-confidence and retaining the constraint of Bayesian rationality, our paper is most closely related to the work of Carrillo and Mariotti [2000], who first showed how information manipulation may serve as a commitment device for time-inconsistent individuals (see also Brocas and Carrillo [1999]). The central role played by memory also relates our model to Mullainathan [2002] and Laibson [2001], although one of its main features is to make recall endogenous.

I. The Demand for Self-Confidence

In most societies, self-confidence is widely regarded as a valuable individual asset. Going back at least to William James, an important strand in psychology has advocated "believing in oneself" as a key to personal success. Today, an enormous "self-help" industry flourishes, a sizable part of which purports to help people improve their self-esteem, shed "learned helplessness" and reap the benefits of "learned optimism."[6] American schools place

6. These last two terms are borrowed from Seligman [1975, 1990].

such a strong emphasis on imbuing children with self-confidence ("doing a great job)") that they are often criticized for giving it preeminence over the transmission of actual knowledge. Hence the general question: why is a positive view of oneself, as opposed to a fully accurate one, seen as such a good thing to have?

Consumption value. A first reason may be that thinking of oneself favorably just makes a person happier: self-image is then simply another argument in the utility function. Indeed, psychologists emphasize the affective benefits of self-esteem as well as the functional ones on which we shall focus. One may also hypothesize that such preferences over beliefs could have been selected for through evolution: the overconfidence that typically results may propel individuals to undertake activities (exploration, foraging, combat) which are more risky than warranted by their private material returns, but confer important external benefits on the species. In subsection V.B we shall explain how a hedonic self-image motive can readily be incorporated into our general framework.

Signaling value. A second explanation may be that believing oneself to be of high ability or morality makes it easier to convince others (rightly or wrongly) that one does have such qualities. Indeed, it is often said that to lie most convincingly a person must believe his own lies. While the idea that people are "transparent" and have trouble misrepresenting their private information may seem unusual in economics, one could easily obtain an instrumental value of self-confidence from a signaling game where those who truly believe in their own abilities face lower costs of representing themselves favorably to others.

Motivation value. The explanation that we emphasize most is that self-confidence is valuable because it improves the individual's motivation to undertake projects and persevere in the pursuit of his goals, in spite of the setbacks and temptations that periodically test his willpower. Morale is universally recognized as key to winning a medal, performing on stage, getting into college, writing a great book, doing innovative research, setting up a firm, losing weight, finding a mate, and so forth. The link between self-confidence and motivation is also pervasive in the psychology literature, from early writers like James [1890] to contemporary ones like Bandura [1977], according to whom "be-

liefs of personal efficacy constitute the key factor of human agency" (see also, e.g., Deci [1975] or Seligman [1990]). The motivation theory also readily extends to economic (nonaltruistic) interactions, explaining why people typically prefer self-confident coworkers, managers, employees, teammates, soldiers, etc., to self-doubting ones; and why they spend substantial time and effort supporting the morale of those with whom they end up being matched.[7]

I.A. The Motivation Problem

> 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*].

Consider a risk-neutral individual with a relevant horizon of three periods: $t = 0,1,2$. At date 0 he selects an action that potentially affects both his flow payoff u_0 and his date 1 information structure.[8] At date 1 he decides whether to undertake a task or project (exert effort, which has disutility cost $c > 0$) or not (exert no effort). With some probability θ, which defines his *ability,* the project will succeed and yield a benefit V at date 2; failure generates no benefit. The individual's beliefs over θ (defining his *self-confidence* or self-esteem) are described by distribution functions $F(\theta)$ at date 0 and $F_1(\theta)$ at date 1. In the intervening period new information may be received, or previous signals forgotten; we shall focus here on the first, more standard case, and turn to memory in Section II. Note that with risk neutrality the mean $\bar{\theta}_1 \equiv \int_0^1 \theta \, dF_1(\theta)$ will be a sufficient statistic for F_1; for brevity, we shall also refer to it as the agent's date 1 self-confidence.

Finally, we assume that the individual's preferences exhibit *time-inconsistency,* due to quasi-hyperbolic discounting. There is indeed considerable experimental and everyday evidence that intertemporal choices exhibit a "salience of the present," in the

7. Note that this last observation cannot readily be accounted for by the "signaling" theory of self-confidence either.

8. The simplest date 0 action is thus the choice of the amount of information that will be available at date 1 (e.g., soliciting feedback, taking a test, keeping or destroying records). Alternatively, this information may be derived from the outcome of some activity pursued for its own sake at date 0 (learning by doing, drinking a lot of wine).

sense that discount rates are much lower at short horizons than at more distant ones.[9] Denoting u_t and $\mathsf{E}_t[\cdot]$ the flow payoffs and expectations at $t = 0,1,2$, the intertemporal utility perceived by the individual as of date 1 is

$$u_1 + \beta\delta\mathsf{E}_1[u_2] = -c + \beta\delta\bar{\theta}_1 V \tag{1}$$

when he undertakes the activity, and 0 when he does not. By contrast, intertemporal utility conditional on the same information set at date 1, but evaluated from the point of view of date 0, is

$$u_0 + \beta\mathsf{E}_0[\delta u_1 + \delta^2 u_2|\bar{\theta}_1] = u_0 + \beta\delta[-c + \delta\bar{\theta}_1 V] \tag{2}$$

if the activity is undertaken at date 1, and u_0 otherwise.[10] Whereas δ is a standard discount factor, β reflects the momentary salience of the present. When $\beta < 1$, the individual at date 0 ("Self 0") is concerned about his date 1 ("Self 1's") excessive preference for the present, or *lack of willpower,* which leads to the underprovision of effort (procrastination). Indeed, Self 1 only exerts effort in the events where $\bar{\theta}_1 > c/\beta\delta V$, whereas, from the point of view of Self 0, it should be undertaken whenever $\bar{\theta}_1 > c/\delta V$. Note that while we focus here on the case where the individual's intrinsic ability θ is unknown, it could equally be the expected payoff in case of success V, 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.

I.B. Confidence Maintenance versus Overconfidence

In an important paper Carrillo and Mariotti [2000] showed that, in the presence of time inconsistency (TI), Blackwell garblings of information may increase the current self's payoff. This result can be usefully applied, and further developed, in our context.

Suppose that, at date 0, our individual can choose between just two information structures for date 1. In the finer one, Self 1 learns his ability θ exactly. In the coarser one, he learns nothing

9. See Ainslie [1992, 2001] for the evidence, and Strotz [1956], Phelps and Pollack [1968], Loewenstein and Prelec [1992], Laibson [1997, 2001], and O'Donoghue and Rabin [1999] for formal models and economic implications.

10. Note that the equality in (2) makes use of the identity $\mathsf{E}_0[\theta|\mathsf{E}_1[\theta] = \bar{\theta}_1] = \bar{\theta}_1$, which holds when—and only when—there is no information loss between dates 0 and 1.

that Self 0 did not know: $F_1(\theta) = F(\theta)$, and hence $\bar{\theta}_1 = \int_0^1 \theta \, dF_1(\theta) \equiv \bar{\theta}_F$. Let us first assume that, in the absence of information, Self 1 will undertake the task: $\bar{\theta}_F > c/\beta\delta V$. The value attached by Self 0 to Self 1's learning the value of θ is therefore $\beta\delta$ times

$$(3) \qquad I_F \equiv \int_{c/\beta\delta V}^{1} (\delta\theta V - c)\, dF(\theta) - (\delta\bar{\theta}_F V - c) = \Gamma_F - \Lambda_F,$$

where

$$(4) \qquad \Gamma_F \equiv \int_{0}^{c/\delta V} (c - \delta\theta V)\, dF(\theta),$$

$$(5) \qquad \Lambda_F \equiv \int_{c/\delta V}^{c/\beta\delta V} (\delta\theta V - c)\, dF(\theta).$$

Γ_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 below $c/\delta V$ but he is unaware of it, and thus inappropriately undertakes or perseveres in the project. Λ_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, $c/\delta V < \theta < c/\beta\delta V$, he will procrastinate at date 1 even though, ex ante, it 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 *confidence maintenance* motive is strong enough ($\Lambda_F > \Gamma_F$), the individual will prefer to remain uninformed: $I_F < 0$. More generally, note that I_F is lower, the lower is β. By contrast, in the absence of time inconsistency ($\beta = 1$) we have $\Lambda_F = 0$, and thus $I_F \geq 0$: in classical decision theory, information is always valuable.

The overconfidence effect calls for more information, confidence maintenance for less. This trade-off 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 nonproductive persis-

tence at 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 turn to the case where $\bar{\theta}_F < c/\beta\delta V$. Since Self 1 now always exerts (weakly) less effort than Self 0 would like him to, information can only help the individual restore his deficient motivation. Indeed,

$$I_F = \int_{c/\beta\delta V}^{1} (\delta\theta V - c)\, dF(\theta) > 0. \tag{6}$$

Moreover, I_F is now *higher,* the lower is β. In such situations the individual will avidly seek feedback on 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 is *not monotonic* with respect to initial self-confidence. Indeed, for someone with confidence so low that $\bar{\theta}_F < c/\beta\delta V$, I_F is always positive and increasing with respect to (stochastic) increases in θ.[11] For an individual with $F(c/\delta V) = 0$ but $F(c/\beta\delta V) < 1$, I_F is always negative. Finally, for a person so self-assured that $F(c/\beta\delta V) = 0$, motivation is not a concern (as if β were equal to 1), but neither is overconfidence: $I_F = \Gamma_F = \Lambda_F = 0$. Therefore, there must exist some intermediate range where I_F first declines and becomes negative, then increases back toward zero.

I.C. What Types of People Are Most Eager to Maintain Their Self-Confidence?

Let us now consider two individuals with different degrees of initial self-confidence, and ask which one is least receptive to information. We denote their prior distributions over abilities as $F(\theta)$ and $G(\theta)$, with densities $f(\theta)$,$g(\theta)$ and means $\bar{\theta}_F$,$\bar{\theta}_G$. To make confidence maintenance meaningful, let $\bar{\theta}_F > \bar{\theta}_G > c/\beta\delta V$. For comparing levels of self-confidence, however, just looking at expected abilities turns out not to be sufficient.

DEFINITION 1. An individual with distribution F over ability θ has higher self-confidence than another one with distribution G if the likelihood ratio $f(\theta)/g(\theta)$ is increasing in θ.

11. Rewrite (6) as $I_F = \int_0^1 1_{[\theta \geq c/\beta\delta V]}(\delta\theta V - c)\, dF(\theta)$, where $1_{[\cdot]}$ denotes the indicator function, and note that the integrand is increasing in θ.

Abstracting for the moment from any cost attached to learning or not learning the true ability, it is easy to see from (3) that $I_F \geq 0$ if and only if

$$\int_0^{c/\beta\delta V} \frac{F(\theta)}{F(c/\beta\delta V)}\, d\theta \geq \left(\frac{1-\beta}{\beta}\right)\left(\frac{c}{\delta V}\right). \tag{7}$$

The monotone likelihood ratio property (MLRP) implies that $F(\theta)/F(c/\beta\delta V) \leq G(\theta)/G(c/\beta\delta V)$ for all $\theta \leq c/\beta\delta V$. Therefore, the left-hand side of (7) is smaller under F than under G, meaning that the person with the more positive self-assessment will accept information about his ability for a smaller set of parameters. Intuitively, he has more to lose from information, and is therefore more insecure.

PROPOSITION 1. If an individual prefers not to receive information in order to preserve his self-confidence, so will anyone with higher initial self-confidence: if $I_G < 0$ for some distribution G over θ, then $I_F < 0$ for any distribution F such that the likelihood ratio f/g is increasing.

I.D. Self-Handicapping

A well-documented and puzzling phenomenon is that people sometimes create obstacles to their own performance.[12] In experiments, subjects with fragile self-confidence opt to take performance-impairing drugs before an intelligence test. In real life, people withhold effort, prepare themselves inadequately, or drink alcohol before undertaking a task. They also set themselves overambitious goals, where they are almost sure to fail. Test anxiety and "choking" under pressure are yet other common examples. Psychologists have long suggested that self-handicapping is often a self-esteem maintenance strategy (instinctive or deliberate), directed both at oneself and at others.[13]

12. See, e.g., Berglas and Jones [1978], Arkin and Baumgardner [1985], Fingarette [1985], or Gilovich [1991].

13. See, e.g., Berglas and Baumeister [1993]. Of course, self-handicapping involves both intrapersonal (self-esteem maintenance) and intrapersonal (self-presentation) motives; our model captures only the former. As Baumeister [1998] notes, "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 then reap a double dividend, as this provides an excuse for poor performance both to themselves and to others.

To examine this question, consider an individual with prior beliefs $F(\theta)$, faced at date zero with a choice between an efficient action that (for simplicity) will fully reveal his ability, and an inefficient, "self-handicapping" one that entails an expected cost $h_0(F) \geq 0$, but is totally uninformative about θ. Assuming that $\bar{\theta}_F > c/\beta\delta V$ as before, equation (7) immediately generalizes to show that he will self-handicap if and only if $-\beta\delta I_F \geq h_0(F)$, or

$$(8) \qquad \left(\frac{1-\beta}{\beta}\right)c - \delta V\left[\int_0^{c/\beta\delta V} \frac{F(\theta)}{F(c/\beta\delta V)}\, d\theta\right] \geq \frac{h_0(F)}{\beta\delta F(c/\beta\delta V)}.$$

Note first that multiplying (8) by $F(c/\beta\delta V)$ yields a decreasing function of β on the left-hand side. Therefore, people who are more concerned about sustaining motivation (more time-inconsistent) are more likely to self-handicap, and will choose to do so when the short-run costs of doing so are not too large. Next, let us compare individuals with different prior beliefs about themselves. As before, those who are initially more self-confident have more to lose from learning about their ability: by the MLRP, the left-hand side of (8) is larger than if F were replaced with G. However, the more self-confident are also *less likely* to receive bad news, and this reduces the return on "investing in ignorance:" the MLRP implies that $F(c/\beta\delta V) \leq G(c/\beta\delta V)$, which tends to make the right-hand side of (8) also larger under F than under G. Thus, in general one cannot conclude whether people with higher or lower self-confidence are more likely to self-handicap.[14] This ambiguity is fundamentally linked to the nonmonotonicity noted earlier for the value of information: while the MLRP ensures that the *sign of* I_F varies monotonically with initial beliefs, the *absolute amount* does not. When self-handicapping costs are relatively small, however—which is often the case in experiments—the "more to lose" effect identified in Proposition 1 will prevail.

It is interesting in this respect to note that psychologists have also not reached a firm conclusion on whether high or low self-confidence people are the most defensive of their egos, although there does seem to be somewhat more evidence in favor of the first hypothesis. Thus, Greenier, Kernis, and Wasschul [1995] contrast "humanistically oriented theories, . . . according to

14. A third consideration is whether the expected cost of self-handicapping rises or falls with initial self-confidence: $h_0(F) \gtrless h_0(G)$. In Bénabou and Tirole [2001] we provide examples of tasks that correspond to each case.

which high self-esteem individuals' 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 display 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."

II. 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*].

We now turn to the supply side of the self-esteem problem. Given that a positive self-assessment may be desirable (whether for motivational, signaling, or hedonic reasons), what are the means through which it can be achieved, or at least pursued?

Wired-in optimism. A first hypothesis could be that evolution has selected for a particular cognitive bias in humans, causing them to systematically and involuntarily underweigh adverse signals about themselves, and overweigh positive ones. This explanation is rather problematic: the extent of overconfidence or overoptimism varies both over time and across tasks, and a great many people actually suffer from underconfidence (the extreme case being depression). Furthermore, individuals often "work" quite hard at defending their self-image when it is threatened, going though elaborate schemes of denial, self-justification, furniture-avoidance, and the like.

Blissful ignorance. When self-confidence is valuable capital, it may be preferable to remain uninformed than to put it at risk by exposing oneself to new information. In particular, as seen in the previous section, ex ante strategic ignorance [Carrillo and Mariotti 2000] or even self-handicapping may help a time-inconsistent individual safeguard his motivation. In a context of hedonic beliefs, workers in a hazardous job may not want to know about the exact risks involved [Akerlof and Dickens 1982].

Self-deception. Most often, the relevant issue is not whether to seek or avoid information ex ante (before knowing what it will turn out to say), but how to deal with the good and especially the bad news concerning one's performances and abilities that life inevitably brings. This is where the mechanisms of defensive denial, repression, self-serving attributions, and the like, so prominently emphasized in psychology, come into play. We shall capture this class of phenomena with a simple game-theoretic model of endogenously selective memory.

II.A. Managing Awareness: The Role of Memory

Psychologists, and before them writers and philosophers, have long documented people's universal tendency to deny, explain away, and selectively forget ego-threatening information. Freudian repression is the most obvious example, but various other forms of *motivated cognition* and *self-deception* figure prominently in contemporary psychology. Thus, a lot of research has confirmed that people tend to recall their successes more than their failures (e.g., Korner [1950] and Mischel, Ebbesen, and Zeiss [1976]), have self-servingly biased recollections of their past performances [Crary 1966] and readily find "evidence" in their personal histories that they possess characteristics which they view (sometimes as the result of experimental manipulation) as correlated with success in professional or personal life [Kunda and Sanitioso 1989; Murray and Holmes 1993]. Similarly, they often engage in "beneffactance," viewing themselves as instrumental for good but not bad outcomes [Zuckerman 1979]. When they commit a bad deed, they reframe the facts to try and convince themselves that it was not so bad ("he deserved it," "the damage was limited"), or attribute the responsibility to others [Snyder 1985].

At the same time, the impossibility of simply choosing the beliefs we like has always stood in the way of a fully consistent theory of self-deception. 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.

Our intertemporal model allows us to unbundle the "self that knows" from the "self that doesn't know," and thereby reconcile the motivation ("hot") and cognition ("cold") aspects of self-decep-

tion within a standard information-theoretic framework. The basic idea is that the individual can, within limits, *affect the probability of remembering* a given piece of data. Under time inconsistency, there is an incentive to try to recall signals that help sustain long-term goals, and forget those that undermine them. This is the motivation part.[15] On the other hand, we maintain the rational inference postulate, so people realize (at least to some extent) that they have a selective memory or attention. This is the cognition part.

ASSUMPTION 1 (MEMORY OR AWARENESS MANAGEMENT). The individual can, at a cost, increase or decrease the probability of remembering an event or its interpretation.

Formally, let $\lambda \in [0,1]$ denote the probability that given information received at date 0 will be recalled or accessed at date 1. We define the natural rate of recall $\lambda_N \in [0,1]$ as that which maximizes the date 0 flow payoff u_0. Increasing or decreasing λ thus 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 $\lambda > \lambda_N$.[16]

Whether it refers to the subconscious or points to the differential accessibility and decay of memories stored in specialized parts of the brain, virtually all modern psychology recognizes that only part of the individual's accumulated stock of information is readily available for conscious, purposive processing and decision-making. Furthermore, the encoding and retrieval process is subject to systematic influences, both internal and external: a)

15. Alternatively, it could arise from an affective or signaling value of self-esteem.

16. By definition, "forgetting" means that an actual *loss of information* (a coarsening of the informational partition) occurs. Thus, if at date 1 the individual does not remember a date 0 signal σ, he also does not recall any other piece of information that is perfectly correlated with σ, such as the costs $M(\lambda)$ incurred in the process of forgetting. This complete forgetting is only a simplified representation of a richer attribution (signal extraction) problem, however. Let d measure the level of an action that affects recall but can also be undertaken for its own sake: amount of wine consumed, time spent with friendly rather than critical people, attention paid to the details of competing information, effort in making, safekeeping, or disposing of physical records, spatial or mental detours around certain potential cues, etc. Choosing d following an event σ leads to a recall probability $\lambda = \Lambda(d)$ and has a direct utility $u_0(d,\epsilon)$, where ϵ is a random taste shock. Later on, the agent may recall the action $d^*(\sigma,\epsilon) \in \arg\max_d\{u_0(d,\epsilon) + \beta\delta E_0[u_1 + \delta u_2|d,\sigma]\}$ that he took, and possibly the associated consequences $u_0(d^*(\sigma,\epsilon),\epsilon)$, but not the particular realization of ϵ that occurred (e.g., how much did I really want to drink wine, or sit next to that person at dinner?). Recalling $d^*(\sigma,\epsilon)$ is thus generally insufficient to fully reconstruct σ, or to separate out within realized utility the cost $M(\Lambda(d^*(\sigma,\epsilon),\epsilon)) \equiv \max_d u_0(d,\epsilon) - u_0(d^*(\sigma,\epsilon),\epsilon)$ that was incurred purely for memory manipulation. This problem remains when the functions u_0, Λ or M, or the distribution of ϵ, depend on σ.

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 remembering the details; b) direct behavioral experience makes the information more accessible in memory, because later on recall is more likely to be activated by situational *cues*.[17]

Such mechanisms seem to be at work in experiments where subjects who are asked 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 [Jones et al. 1981]. This may be due to the fact that they were led to rehearse unfavorable or favorable information about themselves, thus increasing the probability of remembering it later on. Similarly, 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 [Greenwald 1980].

These frictions in the mechanics of memory give the individual some discretion about what data he is more likely to consciously recall later on—thereby opening the door to motivated cognition. Thus, a person who wishes to remember good news and forget the bad can linger over praise or positive feedback, rehearse it periodically, and choose to be more frequently in environments or with people who will remind him of his past successes.[18] Conversely, he can eschew situations that remind him of bad news, tear up the picture of a former lover, 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 can also find ways to discount self-threatening news in the first place. A common such strategy is to seek out information that derogates the informativeness of the initial data [Frey 1981; Gilovich 1991]. After being criticized in a seminar or

17. For evidence and discussions see, e.g., Schacter [1996] and Fazio and Zanna [1981]. Mullainathan [2002] and Laibson [2001] provide models of cue-dependent consumption.

18. Thus, "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 criticisms" [Gilbert and Cooper 1985]. For a discussion of self-presentation strategies and their link with self-enhancement, see Rhodewalt [1986].

referee report, a researcher will look hard for reasons why the commenter has poor taste, a limited understanding of the issues, a vested interest in a competing theory or body of empirical evidence, and so forth. Interpersonal strategies can also be called upon; thus, a verbal fight with one's spouse or someone who criticizes one's work may (consciously or not) serve the purpose of creating a distraction that will impair accurate recollection of the details of the criticism (of course, it has costs as well . . .).

As these examples make clear, 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 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, the process of selective rehearsal afterwards, and again the choice of attention at the time the information is (voluntarily or accidentally) retrieved.[19] We shall therefore use the terms "memory" and "awareness of past information" interchangeably.

ASSUMPTION 2 (METACOGNITION). While the individual can manipulate his conscious self-knowledge, he is aware that incentives exist that result in selective memory.

As illustrated by the opening quotations from Nietzsche and Darwin, if a person has a systematic tendency to forget, distort, or repress certain types of information, he will likely become (or be made) aware of it, and not blindly take at face value the fact that most of what comes to his mind when thinking about his past performances and the feedback he received is good news. Instead, using (some) rational inference, he will realize that what he may have forgotten are not random events.[20] Formally, this *introspection* or skepticism with respect to the reliability of one's own self-knowledge is represented by Bayes' rule, which implies that a person cannot consistently fool himself in the same direction.

19. 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 information.

20. As Gilbert and Cooper [1995] note, "we are all insightful naive psychologists, well aware of human tendencies to be self-serving."

Less sophisticated inference processes lead to similar results, so long as they are not excessively naive (see subsection III.A).

II.B. The Game of Self-Deception

Let the agent receive, at date 0, a signal σ about his ability θ. To make things simple, let σ take only two values: with probability $1 - q$ the agent receives bad news, $\sigma = L$; and with probability q he receives no news at all, $\sigma = \varnothing$. In other words, "no news is good news." Let

$$\theta_L \equiv \mathrm{E}[\theta|\sigma = L] < \mathrm{E}[\theta|\sigma = \varnothing] \equiv \theta_H. \tag{9}$$

Since σ is informative about the return to date 1 effort, the agent's Self 1 would benefit from having this signal. If it is ego-threatening, however, Self 0 may have an interest in suppressing it. The recollection at date 1 of the news σ will be denoted $\hat{\sigma} \in \{\varnothing, L\}$. We assume that memories can be lost but not manufactured ex nihilo, so $\sigma = \varnothing$ always leads to $\hat{\sigma} = \varnothing$. A signal $\sigma = L$, on the other hand, may be forgotten due to natural memory decay or voluntary repression. Let λ denote the probability that bad news will be remembered accurately:

$$\lambda \equiv \Pr[\hat{\sigma} = L|\sigma = L]. \tag{10}$$

As explained earlier, the agent can increase or decrease this probability with respect to its "natural" value $\lambda_N \leq 1$; choosing a recall probability λ involves a "memory cost" $M(\lambda)$. We shall now analyze the equilibrium in several stages.

1. Inference problem of Self 1. Faced with a memory $\hat{\sigma} \in \{L, \varnothing\}$, Self 1 must first assess its credibility. Given that memories cannot be invented, unfavorable ones are always credible. When Self 1 does not recall any adverse signals, on the other hand, he must ask himself whether there was indeed no bad news at date 0, or whether it may have been lost or censored. 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

$$r^* \equiv \Pr[\sigma = \varnothing|\hat{\sigma} = \varnothing; \lambda^*] = \frac{q}{q + (1 - q)(1 - \lambda^*)}. \tag{11}$$

His degree of self-confidence is then

$$\theta(r^*) \equiv r^*\theta_H + (1 - r^*)\theta_L. \tag{12}$$

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 that $\bar{c} > \beta\delta\theta_H > \beta\delta\theta_L > \underline{c}$, which means that at date 1 there is always a positive probability of no effort, and a positive probability of effort.[21]

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 $\mathsf{E}[\theta|\sigma]$ and $\mathsf{E}[\theta|\hat{\sigma}]$. Self 1 only works when the realization of the effort cost is $c < \beta\delta\mathsf{E}[\theta|\hat{\sigma}]$, so Self 0's payoff is

$$\beta\delta \int_0^{\beta\delta\mathsf{E}[\theta|\hat{\sigma}]} (\delta\mathsf{E}[\theta|\sigma] - c)\, d\Phi(c). \tag{13}$$

To build intuition, suppose for a moment that Self 0 could freely and costlessly manipulate Self 1's expectation, $\mathsf{E}[\theta|\hat{\sigma}]$. What beliefs would he choose for a naive Self 1? Maximizing (13), we find that Self 0 would like to set $\mathsf{E}[\theta|\hat{\sigma}]$ equal to $\mathsf{E}[\theta|\sigma]/\beta$. This makes clear how time consistency 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. These two effects are consistent with the common view in psychology that a moderate amount of "positive illusion" about oneself is optimal, but that many people find it quite difficult to strike this desirable balance.

3. Costs and benefits of selective memory or attention. Focusing on the "bad news" case, denote as $U_C(\theta_L|r^*)$ the expected utility of Self 0 (gross of memory-management costs) when the adverse information is successfully forgotten, and as $U_T(\theta_L)$ the corresponding value when it is accurately recalled. The subscripts C and T stand for "censored" and "truth," respectively. Hiding from Self 1 the signal $\sigma = L$ raises his self-confidence from θ_L to $\theta(r^*)$, leading him to exert effort in the additional states of the world where $\beta\delta\theta_L < c < \beta\delta\theta(r^*)$. As with ex ante ignorance, this has both costs and benefits; thus if r^* is high enough that $\beta\theta(r^*) > \theta_L$, the net *gain or loss from self-deception* is

21. In Section I 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 $\delta\theta V/c$.

$$(14)\quad U_C(\theta_L|r^*) - U_T(\theta_L) = \beta\delta\left(\int_{\beta\delta\theta_L}^{\delta\theta_L} (\delta\theta_L - c)\, d\Phi(c) - \int_{\delta\theta_L}^{\beta\delta\theta(r^*)} (c - \delta\theta_L)\, d\Phi(c)\right).$$

The first integral is decreasing in β, becoming zero at $\beta = 1$: it represents the *gain from confidence-building,* which alleviates Self 1's motivation problem. The second integral is increasing in β: it reflects the *loss from overconfidence,* which causes Self 1 to attempt the task in states of the world where even Self 0 would prefer that he abstain. Note that these effects are now endogenous. Thus, the overconfidence cost in (14) is larger, 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 $\beta\theta(r^*) < \theta_L$, the overconfidence effect disappears entirely, but the confidence-building effect is now limited by $\theta(r^*)$.

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

$$(15)\quad \max_{\lambda} \omega\lambda U_T(\theta_L) + (1 - \lambda)U_C(\theta_L|r^*) - M(\lambda)\}.$$

Given the convexity of $M(\lambda)$, the optimum is uniquely determined (given r^*) by the first-order condition, which involves comparing the marginal benefit from self-deception, $U_C(\theta_L|r^*) - U_T(\theta_L)$, with the marginal cost $M'(\lambda)$. Finally, the Bayesian rationality of Self 1 means that he is aware of Self 0's choosing the recall strategy λ opportunistically according to (15), 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 $(\lambda^*, r^*) \in [0,1] \times [q,1]$ that solves (11) and (15), meaning that

i) The recall strategy of Self 0 is optimal, given Self 1's assessment of the reliability of memories.
ii) Self 1 assesses the reliability of memories using Bayes' rule and Self 0's recall strategy.

We shall be interested in two main issues.

1. Nature and multiplicity of equilibria. What modes of self-esteem management are sustainable (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 cognition and 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 his self-confidence and awareness as he sees fit, or prefer a priori to find mechanisms (friends, mates, environments, occupations, etc.) that ensure that he will always be confronted with the truth about himself, no matter how unpleasant it turns out to be?

Because PBE's are related to the solutions $r^* \in [q,1]$ to the nonlinear equation obtained by substituting (11) into the first-order condition for (15), namely

$$(16)\quad \psi(r,\beta) \equiv \beta\delta \int_{\beta\delta\theta_L}^{\beta\delta(r\theta_H+(1-r)\theta_L)} (\delta\theta_L - c)\, d\Phi(c) + M'\left(\frac{1-q/r}{1-q}\right) = 0,$$

we shall use a sequence of simpler cases to demonstrate the main points that emerge from our model. Note, for further reference, that $\psi(r,\beta)$ represents Self 0's (net) *marginal incentive to forget.*

II.C. Costless Memory or Awareness Management

> Repression is automatic forgetting [Henry Laughlin, *The Ego and Its Defenses* 1979].

We first solve the model in the case where the manipulation of memory is costless, $M \equiv 0$. While it does not capture the psychological costs of repression (as opposed to the informational ones), this case already yields several key insights, and is very tractable.[22]

PROPOSITION 2. When $M \equiv 0$, there exist $\underline{\beta}$ and $\bar{\beta}$ in (0,1), $\underline{\beta} < \bar{\beta}$, with the following properties. For low degrees of time inconsistency, $\beta > \bar{\beta}$, the unique equilibrium involves minimum repression ($\lambda^* = 1$); for high degrees, $\beta < \underline{\beta}$, it involves

22. While we assume here 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.

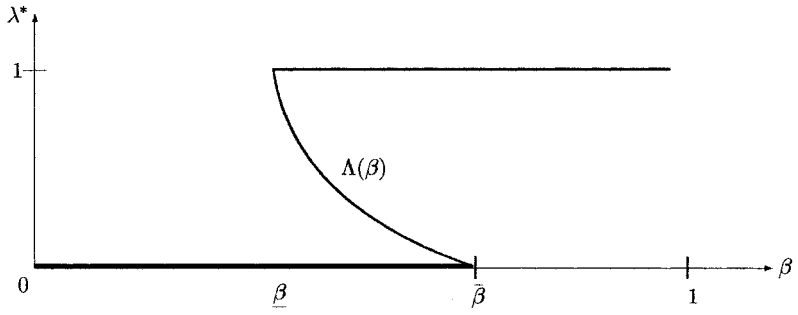

FIGURE I

maximum repression ($\lambda^* = 0$). For intermediate degrees of time inconsistency, $\beta \in [\underline{\beta},\bar{\beta}]$, there are three equilibria, including a partially repressive one: $\lambda^* \in \{0,\Lambda(\beta),1\}$, where $\Lambda(\beta)$ decreases from 1 to 0 as β rises from $\underline{\beta}$ to $\bar{\beta}$.

These results are illustrated on Figure I. The intuition is simple, and apparent from (14). When β is high enough, overconfidence is the dominant concern; therefore, adverse signals are systematically transmitted. Conversely, for low values of β the confidence-building motive dominates, so ego-threatening signals are systematically forgotten. For intermediate values both effects are relevant, allowing multiple equilibria, including one where memory is partially selective. What makes all three equilibria self-fulfilling is precisely the introspection or "metacognition" of the Bayesian individual, who understands that his self-knowledge is subject to opportunistic distortions. The more 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 encodes all news into memory, Self 1 is more likely to be overconfident when he cannot recall any bad signals, and this incites Self 0 to be truthful.

Note that in the censoring equilibrium ($\lambda^* = 0$), none of Self 0's information is ever transmitted to Self 1: $r^* = q$. In the language of communication games, this is a "babbling equilibrium." The mechanism at work here is nonetheless very different from the ex ante suppression of information considered earlier when analyzing self-handicapping, or in Carrillo and Mariotti [2000]. 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.

The last observation to be drawn from Figure I is that, as β rises from 0 to 1, there is necessarily (i.e., for any equilibrium selection) at least one point where λ^* has an upward discontinuity. Small differences in the psychic or material costs of memory management, repression, etc., can thus imply large changes in the selectivity of memory, hence in the variability of self-confidence, and ultimately in performance.

II.D. Costly Memory or Awareness Management

> To break down the renewed assaults of my memory, my imagination effectively labored in the opposite direction [Marcel Proust, *Remembrance of Things Past*].

In this subsection we use specific functional forms to study the problem set up in subsection II.B. The memory cost function is

$$M(\lambda) = a(1 - \ln \lambda) + b(1 - \ln(1 - \lambda)), \tag{17}$$

with $a > 0$ and $b \geq 0$. It is minimized at the "natural" recall rate $\lambda_N = a/(a + b)$, and precludes 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,* namely $U_C(\theta_L|r^*) - U_T(\theta_L)$ in (16), is proportional to a third degree polynomial in r, with either one or three roots in $[q,1]$ (see the Appendix). Therefore, for any (a,b) there are again either one or three equilibria. One can go further, and obtain explicit comparative statics results, by focusing on the simpler case where *recall is costless* but *repression is costly*. The following proposition is illustrated in Figures IIa to IIb.

PROPOSITION 3. Let $b = 0$. A higher degree of time inconsistency or a lower cost of repression increases the scope for memory manipulation, generating partially repressive equilibria and possibly even making perfect recall unsustainable. Formally,

1) For any given β there exist thresholds $\underline{a}$ and $\bar{a}$ with $0 \leq \underline{a} \leq \bar{a}$, and continuous functions $\lambda_1(a),\lambda_2(a)$, respectively, increasing and decreasing in a, such that (i) for $a \in (0,\underline{a})$, the unique

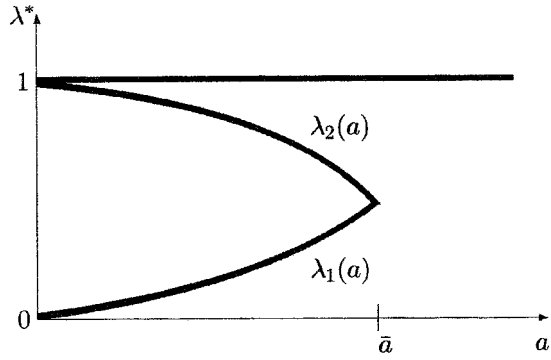

FIGURE IIa
Case $\beta_2 < \beta < \beta_3$

equilibrium corresponds to $\lambda^* = \lambda_1(a)$; (ii) for $a \in (\underline{a},\bar{a})$, there are three equilibria: $\lambda^* \in \{\lambda_1(a),\lambda_2(a),1\}$; (iii) for $a \in (\bar{a},+\infty)$, the unique equilibrium corresponds to $\lambda^* = 1$.

2) There exist critical values $\beta_1 < \beta_2 < \beta_3$ such that (i) for $\beta \geq \beta_3$, $\bar{a} = 0$; (ii) for $\beta \in [\beta_2,\beta_3)$, $\underline{a} = 0 < \bar{a}$, as in Figure IIa; (iii) for $\beta \in (\beta_1,\beta_2)$, $0 < \underline{a} < \bar{a}$, as in Figure IIb; (iv) for $\beta \leq \beta_1$, $0 < \underline{a} = \bar{a}$, as in Figure IIc.

The most representative case is that of Figure IIb, where each of the three ranges $[0,\underline{a}]$, $[\underline{a},\bar{a}]$, and $[\bar{a}, +\infty)$, corresponding, respectively, to high repression, multiplicity, and truthfulness, is nonempty. The effects of a are intuitive; we just note that

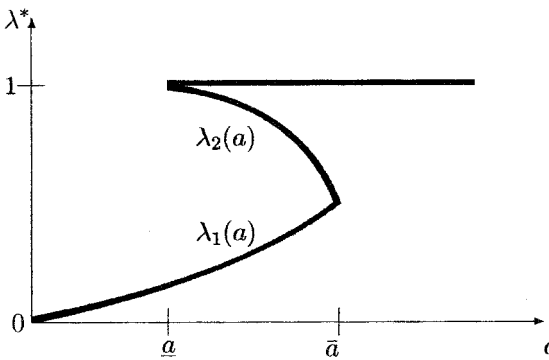

FIGURE IIb
Case $\beta_1 < \beta < \beta_2$

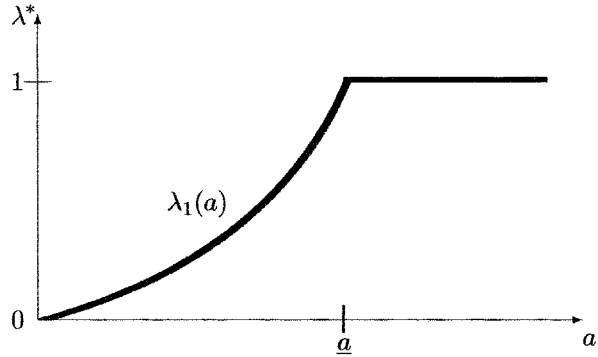

FIGURE IIc
Case $\beta < \beta_1$

small changes in awareness costs can induce large changes in self-esteem and behavior.[23] Interestingly, a lower willpower β, by shifting the equilibrium set toward lower λ's, tends to make the individual incur higher repression costs.

III. BELIEFS AND MAKE-BELIEFS[24]

As discussed earlier, surveys, experiments, and daily observation consistently suggest that most people overestimate their past achievements, abilities, and other desirable traits, both in absolute terms and relative to others (e.g., Weinstein [1980] and Taylor and Brown [1988]). Well-educated, reflective individuals seem to be no exception: as Gilovich [1991] relates, "a survey of college professors found that 94% thought they were better than their average colleague." These findings are often put forward as evidence of pervasive irrationality in human inference.[25] It turns

23. It is also interesting to note that the specification with uniformly distributed costs 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 $\beta\delta(\delta\theta e - e^2/2)$ for Self 0. Thus, in our model, discontinuities in behavior are not predicated on an indivisibility.

24. The terminology of "beliefs and make-beliefs" is borrowed from Ainslie [2001].

25. Without denying the validity of this kind of evidence, we would 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). Second, for every person who is "overconfident" about how great they are (professionally, intellectually, socially,

out, however, that rational self-deception by Bayesian agents can help account for most people holding biased, self-serving beliefs, which in turn have aggregate effects.

III.A. Optimistic and Pessimistic Biases

Continuing to work with our awareness-management model, let us compare the cross-sectional distributions of true and self-perceived abilities.[26] In a large population, a proportion $1 - q$ of individuals are of low ability θ_L, having received a negative signal, $\sigma = L$. The remaining q, having received $\sigma = \varnothing$, have high ability θ_H. Average ability is $q\theta_H + (1 - q)\theta_L = \theta(q)$; we 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 $\lambda^* \in (0,1)$.[27] As before, let r^* denote the corresponding reliability of memory, given by (11). 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 or forgetting are low enough (e.g., a small a in Figure IIc), one can easily have $(1 - q)(1 - \lambda^*) > 1/2$, and 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.*[28] Adding those who had truly received good news ($\sigma = \varnothing$), the fraction of the population who think they are better than average is even larger, namely $1 - \lambda^*(1 - q)$. The remaining minority think,

---

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 confidant, counselor, or therapist. These could even be the same people at different points in time.

26. The interesting distinction is between self-perceived ability $E[\theta|\hat{\sigma}]$ and objectively assessed ability $E[\theta|\sigma]$, rather than between $E[\theta|\sigma]$ and the individual's true θ which it may measure only imperfectly. To simplify the exposition, we shall therefore assume in this section that σ is perfectly informative about θ; i.e., $\theta = E[\theta|\sigma] \in \{\theta_L,\theta_H\}$. Alternatively, one could just read "objectively assessed ability" wherever "ability" appears.

27. Either this is the unique equilibrium, as in Figure IIc, or else we focus on a symmetric situation for simplicity.

28. Note that these statements (like most experimental data) are about the agent's perception of his rank in the distribution of true abilities—not in the distribution of self-assessments which, as a Bayesian, he realizes are generally overoptimistic.

correctly, that they are worse than average; as a result they have low motivation and are unlikely to undertake challenging tasks. They fit the experimental findings of depressed people as "sadder but wiser" realists, compared with their nondepressed counterparts who are much more likely to exhibit self-serving delusions [Alloy and Abrahamson 1979].

As seen above, Bayes' law does not constrain the skewness in the distribution of biases:[29] it only requires that the *average* bias across the $(1 - q)(1 - \lambda^*)$ optimists and the q pessimists be zero: indeed, $(1 - q)(1 - \lambda^*)r^* - q(1 - r^*) = 0$ by (11). In other words, Bayesian rationality only imposes a trade-off between the relative proportions of overconfident versus underconfident agents in the population, and their respective degrees of over- or underconfidence. 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.

III.B. To Bayes or Not to Bayes?

Having shown that even rational agents can deceive themselves most of the time (albeit not all the time), we nonetheless recognize that it may be more realistic to view people as *imperfect Bayesians* who do not fully internalize the fact that their recollections may be self-serving. At the other extreme, taking beliefs as completely naive would be even more implausible. As argued in subsection II.A (also recall Nietzsche and Darwin), 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, when practiced, always optimal (ex ante). With even *some* of this metacognition, self-deception becomes a much more subtle and complex endeavor.

In Bénabou and Tirole [1999] we relax Bayesian rationality and allow the agent at date 1 to remain unaware not just of *what*

29. This was first pointed out by Carrillo and Mariotti [2000] for strategic ignorance, and is a feature that our model also shares with those of Brocas and Carrillo [1999] and Köszegi [1999].

he may have forgotten, but also of the fact *that* he forgets. Self 1's assessment of the reliability of a recollection $\hat{\sigma} = \varnothing$ is thus modified to

$$(18) \qquad r_{\pi}(\lambda) \equiv \Pr[\sigma = \varnothing | \hat{\sigma} = \varnothing; \lambda] = \frac{q}{q + \pi(1-q)(1-\lambda)},$$

where λ is the actual recall strategy and $\pi \in [0,1]$ parameterizes cognitive sophistication, ranging from complete naivete ($r_0(\lambda) \equiv 1$) to full rationality ($r_1(\lambda) \equiv r^*(\lambda)$). As long as π is above a critical threshold, meaning that the individual's self-conception is not too unresponsive to his actual pattern of behavior, all the Bayesian model's results on multiplicity and welfare rankings of personal equilibria go through. Thus, for the purpose of understanding self-deception and overoptimism, the explanatory power gained by departing from rational inference is rather limited (perception biases need no longer sum to zero), whereas much can be lost if the departure is too drastic: without sufficient introspection, one cannot account for "self-traps" or self-doubt.

IV. 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] and Seligman [1990]), pointing to studies which tend to show that a moderate dose of "positive illusions" has significant affective and functional benefits.[30] On the other side are skeptics and outright critics (e.g., Baumeister [1998] and Swann [1996]), who see instead a lack of convincing evidence, and

30. There is of course a huge industry based on that premise, with countless web sites devoted to "self-esteem," 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.

point to the dangers of overconfidence as well as the loss of standards that results when negative feedback is systematically withheld 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 (11)). With probability $1 - q$, Self 0 receives bad news, which he then forgets with probability $1 - \lambda^*$; the resulting expected payoff is $\lambda^* U_T(\theta_L) + (1 - \lambda^*)U_C(\theta_L|r^*) - M(\lambda^*)$. With probability q the news is 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. Indeed, the payoff to Self 0 following genuinely "good news" is only

$$U_T(\theta_H|r^*) = \beta\delta \int_0^{\beta\delta\theta(r^*)} (\delta\theta_H - c)\, d\Phi(c), \tag{19}$$

which is clearly less than $U_T(\theta_H|1)$ 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 with a situation where Self 0 always truthfully records all events into memory. Like a ruler whose entourage dares not bring him bad news, or a child whose parents praise him indiscriminately, an individual with some understanding of the self-serving tendency in his attention or memory can never be sure that he really "did great," even in instances where this was actually true.

Averaging over good and bad news, the agent's ex ante welfare in equilibrium equals

$$\begin{aligned} W(\lambda^*,r^*) \equiv\ & qU_T(\theta_H|r^*) \\ & + (1 - q)[\lambda^* U_T(\theta_L) + (1 - \lambda^*)U_C(\theta_L|r^*) - M(\lambda^*)]. \end{aligned} \tag{20}$$

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 $\Delta W(\lambda^*,r^*) \equiv W(\lambda^*,r^*) - W(1,1)$, we have

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

or, equivalently,

$$\Delta W(\lambda^*,r^*) = (1-q)\Bigg((1-\lambda^*)\int_{\beta\delta\theta_L}^{\beta\delta\theta(r^*)}(\delta\theta_L - c)\,d\Phi(c) - M(\lambda^*) + M(1)\Bigg) - q\int_{\beta\delta\theta(r^*)}^{\beta\delta\theta_H}(\delta\theta_H - c)\,d\Phi(c). \tag{21}$$

The first expression describes the net *gain from forgetting bad news;* the second one the *loss from disbelieving good news.* While the individual is better motivated or even overmotivated following a negative signal about his ability, he may actually be undermotivated following a good signal. A few general results can be immediately observed.

First, if memory manipulation is costless ($M \equiv 0$), then a partial recall (mixed strategy) equilibrium can never be better than perfect recall. Indeed, in such an equilibrium the gain from hiding bad news is zero ($U_C(\theta_L|r^*) = U_T(\theta_L)$) 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 no longer applies, as the term in large brackets in (21) becomes $M(1) - M(\lambda^*) - (1-\lambda^*)M'(\lambda^*) > 0$, by the convexity of M. Similarly, when systematic denial ($\lambda^* = 0$) is an equilibrium, it generates a positive "surplus" in the event of bad news: $U_C(\theta_L|q) - U_T(\theta_L) > M'(0) \geq 0$. How does this gain compare with the loss from self-distrust in the good-news state? As seen from (21), the key intuition involves the likelihood of cost realizations sufficiently high to discourage effort in the absence of adverse recollections ($\hat{\sigma} = \varnothing$). 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 cost density $\varphi(c)$ decreases fast enough,

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

then ex ante welfare is higher if all bad news is censored from memory than if it is always recalled. If the inequality is

reversed, so is the welfare ranking. For a given cost distribution φ, self-deception is thus more likely to be beneficial for a less time-consistent individual.

Note that even the second result was far from obvious a priori, since both the gain and the loss in (21) decrease with β: in equilibrium, memory manipulation tends to alleviate procrastination when $\sigma = L$, but worsen it when $\sigma = \varnothing$.[31] To summarize, we have shown that

1) When the tasks one faces are very difficult and one's willpower is not that strong, a strategy of active self-esteem maintenance, "looking on the bright side," avoiding "negative" thoughts and people, etc., as advocated in numerous "self-help" books, can indeed pay off.
2) When the typical task is likely to be only moderately challenging, and time inconsistency is relatively mild, one can only lose by playing such games with oneself, and it would be better to always "be honest with yourself" and "accept who you are."

It is important to note that in the second case, the individual *may still play* such denial games, even though self-honesty would be better. First, he could be trapped in an inferior equilibrium. Second, motivated cognition may be the *only* equilibrium, yet still result in lower welfare than if the individual could commit to never try to fool himself.[32] A couple of examples will help make these results more concrete.

a) With a uniform density on $[0,\bar{c}]$ $(\bar{c} > \beta\delta\theta_H)$, self-deception is *always harmful* compared with truth-telling. This applies whether both $\lambda^* = 0$ and $\lambda^* = 1$ are in the equilibrium set, or only one of them. (See Proposition 2 and Figure I.) It also applies, a fortiori, when repression is costly.
b) Conversely, self-deception is *always beneficial* when $\varphi(c) =)\, c^{-n}$ on $[\underline{c},+\infty)$, with $0 < \underline{c} < \beta\delta\theta_L$,) chosen so

31. The 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 $\theta(0) = q$). Only when time inconsistency is strong enough can this value can be negative.

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

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 if $M \equiv 0$.

c) Finally, we provide in the Appendix a simple example where both $\lambda^* = 0$ and $\lambda^* = 1$ coexist as equilibria, 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, 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 $\lambda < \lambda_N$).[33] This reinterpretation requires minor modifications to (21), but the main conclusions remain unaltered.[34]

The potential multiplicity of equilibria in our model raises the issue of coordination among the individual's temporal selves. Observe that Self 1 always values information about the productivity of his own efforts, and therefore always ranks equilibria (or commitment outcomes) in order of decreasing λ's. When Self 0 also prefers the $\lambda^* = 1$ solution, it is plausible (we are agnostic on this point) that the individual will find ways to coordinate on this Pareto superior outcome. When some repression (any $\lambda^* < 1$) is ex ante valuable, however, there is no longer any clearly natural selection rule. In either case, our main welfare conclusions remain unchanged even if one assumes that Self 0 always manages to select his preferred equilibrium. First, for some range of β or a, $\lambda^* = 1$ ceases to be an equilibrium even though its still maximizes ex ante welfare. Thus, once again, the individual is trapped in a harmful pattern of systematic denial. Conversely, for relatively high values of a the only equilibrium may be $\lambda^* = 1$ (more generally, a high λ^*), even though the individual would, ex ante, be better off it he could manage to repress bad news more easily.

Interestingly, our multiplicity and welfare results provide a role for parents, friends, therapists, and other benevolent outside parties to help an individual escape the "self-traps" [Swann 1996]

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

34. Term $M(1)$ in (20)–(21) is simply replaced by $\beta^{-1}\delta^{-]}\bar{M}/(1 - q) + m(1)$, where $\bar{M}$ is the up-front cost of the commitment mechanism, $-] < 0$ 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 $\lambda < 1$).

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 cognitive rules. They may also offer a form of informational commitment, serving as the repositories of facts and feelings which the individual realizes that he 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 $M(\lambda)$ (and hence the set of feasible equilibria), whether through their own feedback and questioning, or by teaching him certain cue-management techniques. Indeed, much of modern cognitive therapy aims at changing people's self-image through selective recollection and rehearsal of events, self-serving attributions about success and adversity, or conversely helping them "see through" harmful self-delusions.

V. Variants and Extensions

V.A. Defensive Pessimism

While people are most often concerned with enhancing and protecting their self-esteem, there are also many instances where they seek to minimize their achievements, or convince themselves that the task at hand will be difficult rather than easy. A student preparing for exams may thus discount his previous good grades as attributable to luck or lack of difficulty. A young researcher may understate the value of his prior achievements, compared with what will be required to obtain tenure. 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.

Such behavior, termed "defensive pessimism" by psychologists, can be captured with a very simple variant of our model. 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 is now a *substitute* rather than a complement to effort in generating future payoffs. This gives the agent an incentive to discount, ignore, and otherwise repress signals of *high* ability, as these would increase the temptation to "coast" or "slack off."

Substitutability may arise directly in the performance "production function" which, instead of the multiplicative form

$\pi(e,\theta) = \theta e$ that we assumed, could be of the form $\pi(e,\theta)$ with $\pi_{e\theta} < 0$. More interestingly, it will typically occur when the *reward* for performance is of a "pass-fail" nature: graduating from school, making a sale, being hired or fired (tenure, partnership), proposing marriage, etc. To see this, let performance remain multiplicative in ability and effort: $\pi(\theta,e,\epsilon) = \epsilon\theta e$, where ϵ is a random shock with cumulative distribution $H(\epsilon)$. The payoff V, however, is now conditional on performance exceeding a cutoff level $\bar{\pi}$. Self 1's utility function is thus

$$\beta\delta V \Pr[\epsilon\theta e \geq \bar{\pi}] - ce = \beta\delta V(1 - H(\bar{\pi}/\theta e)) - ce. \tag{22}$$

It is easily verified that if the density $h = H'$ is such that $xh'(x)/h(x) > -1$ on the relevant range of $x \equiv \bar{\pi}/\theta e$, the optimal effort is decreasing in θ. Note that these results yield *testable* predictions: by comparing subjects' confidence-maintenance behavior across experiments (or careers) where payoffs are complements and substitutes, one should be able to distinguish between the motivation-based theory of self-confidence and the hedonic or signaling alternatives.

An even simpler form of defensive pessimism arises in situations where the action subject to time inconsistency is such that *the benefits precede the costs*. One can think of the trade-off between the immediate pleasure of smoking, drinking, spending freely, etc., and the long-term, large but uncertain costs of such behaviors. Suppose, for instance, that at date 1 the decision is to consume or not consume. The first option yields utility b, but with probability ω entails a cost C at date 2; the second option yields zero at both dates. Clearly, if we define "effort" $e \equiv 1 - x$ as abstinence from consumption, $c = b$ as its (opportunity) cost and ωC as its expected long-term payoff, we see that this problem fits exactly with our model. Thus, to counteract his tendency toward short-term gratification, the agent will try to maintain beliefs that ω and C are high.[35] Note that these variables are complements to $e = 1 - x$ in his utility function. If we had framed the problem in terms of uncertainty over the probability of being immune to the health risks (say) from tobacco or alcohol, this probability $1 - \omega$ would be a substitute with e, so the agent would like to understate it to his future selves. Whether costs precede or

35. This is the kind of example to which Carrillo and Mariotti [2000] apply their model of strategic ignorance, pointing to studies that suggest that most people actually overestimate the health risks from smoking. More generally, the role of the timing of costs and benefits is emphasized in Brocas and Carrillo [2000].

follow benefits thus simply amounts to a relabeling of variables. The only thing that matters for the direction in which the agent would like to manipulate his beliefs concerning a variable is its cross-derivative with the decision variable that is being set inefficiently low due to time inconsistency.

V.B. Self-Esteem as a Consumption Good

We have until now emphasized the value of self-confidence for personal motivation. This approach provides an explanation of both *why* and *how much* people care about their self-image: its value arises endogenously from fundamental preferences, technological constraints, and the structure of incentives. As explained earlier, the motivation theory also readily extends to social interactions.

This functional view of self-esteem, while pervasive in psychology, is by no means the only one (e.g., Baumeister [1998]). As discussed earlier, a common and complementary view involves purely affective concerns: people just *like* to think of themselves as good, able, generous, attractive, and conversely find it painful to contemplate their failures and shortcomings. 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, ultrarational 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 utility in Weinberg [1999] and Köszegi [1999], as well as in some interpretations which Akerlof and Dickens [1982] offered for their model of dissonance reduction. The trade-off between the costs and benefits of information can then be modeled by positing preferences of the form,

$$\mathsf{E}[\max\{\hat{\theta}V - c, 0\} + u(\hat{\theta})], \tag{23}$$

where $\hat{\theta}$ denotes the individual's self-perceived ability (expected

probability of success) at the time of the effort decision.[36] The first term always generates a demand for accurate information, to improve decision-making. Suppose for now that the hedonic valuation $u(\hat{\theta})$ is increasing and concave; these properties, respectively, imply a positive demand for self-esteem, and risk aversion with respect to self-relevant signals. The individual may then, once again, avoid free information or engage in self-handicapping.[37] Similarly, all our results based on memory management on the supply side carry over to this case. Thus, "positive thinking" and similar self-deception strategies may be pursued even though they are 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, again, one of *externalities across information states*: having only good news is not such a great boost to self-esteem once the agent realizes that he would have had reasons to censor any bad news that might have been received.

Unfortunately, psychology provides little guidance on what the appropriate shape of the hedonic preference function should be (by contrast, there is ample evidence of people's general bias toward short-term rewards, tendency to procrastinate, etc.). It is thus equally likely that $u(\hat{\theta})$ is convex, at least over some range; in such cases the individual will be an avid information-seeker, choosing tasks that are excessively hard or risky but very informative, as a way of gambling for (self-) resurrection. Even monotonicity may not be taken for granted, since psychologists have documented both optimism and defensive pessimism. The latter, whether originating from motivation concerns or hedonic ones (lowering one's expectations of performance because surprise

36. A more general formulation, encompassing both affective and instrumental concerns, would be $\mathsf{E}_0[\max\{\mathsf{E}_1[\theta V - c],0\} + J(F_1)]$, where $F_1(\theta)$ denotes the agent's date 0 and date 1 subjective probability distribution over his true ability. The functional $J(\cdot)$ represents either an exogenous hedonic utility, or an endogenous value function capturing the instrumental value of beliefs for self-motivation or self-presentation (signaling) purposes. In our model $J(F_1)$ is easily computed, and related to β.

37. 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. Caplin and Leahy [2001] 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 or, on the contrary, seek out information and situations that increase the stakes.

sharpens both the sweetness of success and the bitterness of defeat), requires that $u(\hat{\theta})$ be sometimes decreasing.

VI. Conclusion

Building on a number of themes from cognitive and social psychology, we proposed in this paper a general economic model of why people value their self-image, and of how they seek to enhance or preserve it through a variety of seemingly irrational behaviors—from handicapping their own performance to practicing self-deception through selective memory or awareness management.

This general framework can be enriched in many directions. On the motivation side, we noted earlier that anyone with a vested interest in an individual's success (or failure) has incentives to manipulate the latter's self-perception. Thus, in principal-agent relationships or bargaining situations, the management of self-confidence will matter even when everyone is fully time-consistent. These issues are explored in Bénabou and Tirole [2001], where we examine the provision of incentives by informed principals (parents, teachers, managers) in educational and workplace environments. Because offering rewards for performance may signal low trust in the abilities of the agent (child, student, worker) or in his suitability to the task, such extrinsic motivators may have only a limited impact on his current performance, and undermine his intrinsic motivation for similar tasks in the future—as stressed by psychologists.

Another interesting direction is to further explore the rich set of behavioral implications that arise from the interaction of imperfect willpower and imperfect memory. Thus, in Bénabou and Tirole [2002] we develop a model of self-reputation over one's willpower that can account for the "personal rules" (diets, exercise regimens, resolutions, moral or religious precepts, etc.) through which people attempt to achieve self-discipline. The sustainability of rule-based behavior is shown to depend on the effectiveness of the individual's *self-monitoring* (recalling past lapses and their proper interpretation), which may be subject to opportunistic distortions of memory or inference of the type studied in the present paper. The model also helps explain why people may sometimes adopt excessively "legalistic" rules that result in compulsive behavior such as miserliness, workaholism, or anorexia.

APPENDIX

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

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

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 $\}(r,B(r)) = 0$ and

i) $\}(r,\beta) > 0$ for all $\beta < B(r)$, while $\}(r,\beta) < 0$ 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 $\}(r,\beta) > 0$ for $\beta \in [0,\theta_L/\theta(r)]$, while $\}(r,1) < 0$. Moreover, for all $\beta > \theta_L/\theta(r)$, we have

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

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

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

so $\}(r,B(r))/r < 0$ 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 $\}(r,\beta) < 0$. 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 $\}(r,\beta) > 0$. 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 $\}(R(\beta),\beta) = 0$. Moreover, R is decreasing and for any $r \in (q,1)$, $\}(r,\beta)$

has the sign of $R(\beta) - r$. Therefore, the only equilibrium with $r < R(\beta)$ is $r = q$ (or $\lambda = 0$), and the only equilibrium with $r > R(\beta)$ is $r = 1$ (or $\lambda = 1$). Finally, $r = R(\beta)$ is also 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. ■

Proof of Proposition 3. We shall solve for equilibria in terms of the reliability of memory, r^*; the recall strategy λ^* is then obtained by inverting (11). With the assumptions of the proposition, the incentive to forget, given by (16), equals

$$\psi(r,\beta) = r(\Delta\theta)\left(\frac{\beta^2\delta^3}{\bar{c}}\right)\left((1-\beta)\theta_L - \frac{\beta r}{2}(\Delta\theta)\right)r - ar\left(\frac{1-q}{r-q}\right) + br\left(\frac{1-q}{q(1-r)}\right), \tag{A.4}$$

where $\Delta\theta \equiv \theta_H - \theta_L$. Defining, for all $\beta \in [0,1]$,

$$R(\beta) \equiv \left(\frac{1-\beta}{\beta}\right)\frac{2\theta_L}{\Delta\theta}, \tag{A.5}$$

$$\Omega(\beta) \equiv (\Delta\theta)^2\left(\frac{\beta^3\delta^3}{2\bar{c}}\right)\left(\frac{q}{1-q}\right). \tag{A.6}$$

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

$$P(r,\beta) \equiv \Omega(\beta)(R(\beta) - r)(r - q) \geq aq - b\left(\frac{r-q}{1-r}\right). \tag{A.7}$$

Multiplying by $(1 - r)$ shows that the sign of $\psi(r,\beta)$ is that of a third-degree polynomial in r, with $\lim_{r \to q} \psi(r,\beta) = -\infty$ since $a > 0$, and $\lim_{r \to 1} \psi(r,\beta) = +\infty$ when $b > 0$. Thus, for a given β there are either one or three solutions to $\psi(r,\beta) = 0$ in $[q,1]$, i.e., one or three equilibria.

Let us now specialize (A.7) to the case where remembering is costless but forgetting or repressing is 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. We shall denote $\bar{a} \equiv q^{-1} \max\{\max_{r \in [q,1]} P(r,\beta), 0\}$ and $\underline{a} \equiv q^{-1} \max\{P(1,\beta),0\} \leq \bar{a}$. 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 $r = 1$. Moreover, $\underline{a} = \bar{a} = 0$.

2) For $q < R(\beta) \equiv \beta_3$ the polynomial $P(r,\beta)$ is positive on $r \in [q,R(\beta)]$, implying $\bar{a} > 0$, and negative outside.
   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. One can associate with these two functions, $\lambda_1(a)$ and $\lambda_2(a)$, by inverting (11). 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 third (and only other) equilibrium is $r = 1$. Furthermore, $\underline{a} = 0$.
      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 $\underline{a} = P(1,\beta) > 0$. 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 $a < \underline{a} = P(1,\beta) = \bar{a}$. It is $r = 1$ whenever $a \geq \underline{a}$. ■

Proof of Proposition 4. Setting $\lambda^* = 0$, $r^* = q$ and $M \equiv 0$ in (21) yields

$$\Delta W(0,q) = (1 - q) \int_{\beta\delta\theta_L}^{\beta\delta\theta(q)} (\delta\theta_L - c)\, d\Phi(c)$$

$$- q \int_{\beta\delta\theta(q)}^{\beta\delta\theta_H} (\delta\theta_H - c)\, d\Phi(c)$$

$$= q \int_0^{\beta\delta\theta(q)} (\delta\theta_H - c)\, d\Phi(c) + (1-q) \int_0^{\beta\delta\theta(q)} \times (\delta\theta_L - c)\, d\Phi(c) - q \int_0^{\beta\delta\theta_H} (\delta\theta_H - c)\, d\Phi(c) - (1-q) \int_0^{\beta\delta\theta_L} (\delta\theta_L - c)\, d\Phi(c)$$

$$= \int_0^{\beta\delta\theta(q)} [\delta(q\theta_H + (1-q)\theta_L) - c]\, d\Phi(c) - q \int_0^{\beta\delta\theta_H} (\delta\theta_H - c)\, d\Phi(c) - (1-q) \int_0^{\beta\delta\theta_L} (\delta\theta_L - c)\, d\Phi(c).$$

Defining the function $\approx(Z,\beta) \equiv \int_0^Z (Z - \beta c)\, d\Phi(c)$, we can then write

$$\Delta W(0,q) = \beta^{-1}[\approx(\beta\delta(q\theta_H + (1-q)\theta_L),\beta) - q\approx(\beta\delta\theta_H,\beta) - (1-q)\approx(\beta\delta\theta_L,\beta)]. \tag{A.8}$$

Clearly, $\Delta W(0,q) > 0$ when $\approx$ is concave in Z, and $\Delta W(0,q) < 0$ when it is convex. Indeed, $\beta\delta\Delta W(0,q)$ is (minus) the ex ante value of information, i.e., of always knowing the true $\mathsf{E}[\theta|\sigma]$ rather than have only the uninformed prior or posterior $\theta(q)$. The proposition immediately follows from the fact that $^2\approx(Z,\beta)/\, Z^2 = (2 - \beta)\varphi(Z) + (1 - \beta)Z\varphi'(Z)$. ■

Welfare Rankings of Multiple Equilibria. We construct here a simple example where $\lambda^* = 0$ and $\lambda^* = 1$ coexist as equilibria, and where *either one* can 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

$$\frac{\pi}{1-\pi} < \frac{\bar{c} - \delta\theta_L}{\delta\theta_L - \underline{c}} \equiv \Sigma. \tag{A.9}$$

With this condition both $\lambda^* = 0$ and $\lambda^* = 1$ are equilibria (with a mixed-strategy one in between, which is always dominated by $\lambda^* = 1$ since $M \equiv 0$), and $\lambda^* = 0$ yields higher welfare when

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

or

$$\frac{\pi}{1-\pi} > \left(\frac{q}{1-q}\right)\left(\frac{\delta\theta_H - \bar{c}}{\delta\theta_L - \underline{c}}\right) \equiv \Sigma'. \tag{A.10}$$

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

PRINCETON UNIVERSITY, NATIONAL BUREAU OF ECONOMIC RESEARCH, AND CENTRE FOR ECONOMIC POLICY RESEARCH
INSTITUT D'ECONOMIE INDUSTRIELLE, GREMAQ/CNRS, CERAS/CNRS, ECOLE DES HAUTES ETUDES EN SCIENCES SOCIALES, AND MASSACHUSETTS INSTITUTE OF TECHNOLOGY

REFERENCES

Ainslie, G., *Picoeconomics: The Strategic Interaction of Successive Motivational States within the Person* (*Studies in Rationality and Social Change*) (Cambridge, UK: Cambridge University Press, 1992).
——, *Breakdown of Will* (Cambridge, UK: Cambridge University Press, 2001).
Akerlof, G., and W. Dickens, "The Economic Consequences of Cognitive Dissonance," *American Economic Review,* LXXII (1982), 307–319.
Alloy, L. T., and L. Abrahamson, "Judgement of Contingency in Depressed and Nondepressed Students: Sadder but Wiser?" *Journal of Experimental Psychology: General,* CVIII (1979), 441–485.
Arkin, R. M., and A. H. Baumgardner, "Self-Handicapping," in *Attribution: Basic Issues and Applications,* J. Harvey and G. Weary, eds. (New York: Academic Press, 1985).
Bandura, A., *Self Efficacy: The Exercise of Control* (New York: W. H. Freeman Company, 1977).
Baumeister, R., "The Self," in *The Handbook of Social Psychology,* D. Gilbert, S. Fiske, and G. Lindzey, eds. (Boston, MA: McGraw-Hill, 1998).
Bénabou, R., and J. Tirole, "Self-Confidence: Intrapersonal Strategies," IDEI mimeo, June 1999.
Bénabou, R., and J. Tirole, "Intrinsic and Extrinsic Motivation," Princeton University, mimeo, December 2001.
Bénabou, R., and J. Tirole, "Willpower and Personal Rules," CEPR Discussion Paper No. 3143, January 2002.
Berglas, S., and E. Jones, "Drug Choice as a Self-Handicapping Strategy in

Response to Noncontingent Success," *Journal of Personality and Social Psychology,* XXXVI (1978), 405–417.
Brocas, I., and J. Carrillo, "Entry Mistakes, Entrepreneurial Boldness and Optimism," ULB-ECARE mimeo, June 1999.
Brocas, I., and J. Carrillo, "The Value of Information When Preferences Are Dynamically Inconsistent," *European Economic Review,* XLIV (2000), 1104–1115.
Caplin, A., and J. Leahy, "Psychological Expected Utility Theory and Anticipatory Feelings," *Quarterly Journal of Economics,* CXVI (2001), 55–79.
Carrillo, J., and T. Mariotti, "Strategic Ignorance as a Self-Disciplining Device," *Review of Economic Studies,* LXVI (2000), 529–544.
Crary, W. G., "Reactions to Incongruent Self-Experiments," *Journal of Consulting Psychology,* XXX (1966), 246–252.
Darwin, F., *The Life and Letters of Charles Darwin,* Edited by his son Francis Darwin (New York: D. Appleton and Co., 1898).
Deci, E., *Intrinsic Motivation* (New York: Plenum Press, 1975).
Elster, J., "Motivated Belief Formation," Columbia University, mimeo, June 1999.
Fazio, R., and M. Zanna, "Direct Evidence and Attitude-Behavior Consistency," in L. Berkowitz, ed., *Advances in Experimental Social Psychology,* Vol. 14 (New York: Academic Press, 1981).
Festinger, L., "A Theory of Social Comparison Processes," *Human Relations,* VII (1954), 117–140.
Fingarette, H., "Alcoholism and Self-Deception," in *Self-Deception and Self-Understanding,* M. Martin, ed. (Lawrence, KS: University Press of Kansas, 1985).
Freud, S., *A General Introduction to Psychoanalysis* (New York: Garden City Publishing Co., 1938).
Frey, D., "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,* XVII (1981), 42–50.
Gilbert, D., and J. Cooper, "Social Psychological Strategies of Self-Deception," in *Self-Deception and Self-Understanding,* M. Martin, ed. (Lawrence, KS: University Press of Kansas, 1985).
Gilovich, T., *How We Know What Isn't So* (New York: Free Press, 1991).
Greenier, K., M. Kernis, and S. Wasschul, "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, 1995).
Greenwald, A., "The Totalitarian Ego: Fabrication and Revision of Personal History," *American Psychology,* XXXV (1980), 603–613.
Gur, R., and H. Sackeim, "Self-Deception: A Concept in Search of a Phenomenon," *Journal of Personality and Social Psychology,* XXXVII (1979), 147–169.
Heider, F., *The Psychology of Interpersonal Relations* (New York: Wiley, 1958).
James, W., *The Principles of Psychology* (Cleveland, OH: World Publishing, 1890).
Jones, E., F. Rhodewalt, S. Berglas, and J. Skelton, "Effects of Strategic Self-Presentation on Subsequent Self-Esteem," *Journal of Personality and Social Psychology,* XL (1981), 407–421.
Kolditz, T., and R. Arkin, "An Impression-Management Interpretation of the Self-Handicapping Strategy," *Journal of Personality and Social Psychology,* XLIII (1982), 492–502.
Korner, I., *Experimental Investigation of Some Aspects of the Problem of Repression: Repressive Forgetting,* Contributions to Education, No. 970 (New York: Bureau of Publications, Teachers' College, Columbia University, 1950).
Köszegi, B., "Self-Image and Economic Behavior," MIT, mimeo, October 1999.
Kunda, Z., and R. Sanitioso, "Motivated Changes in the Self-Concept," *Journal of Personality and Social Psychology,* LXI (1989), 884–897.
Laibson, D., "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics,* CXII (1997), 443–478.
——, "A Cue-Theory of Consumption," *Quarterly Journal of Economics,* CXVI (2001), 81–119.
Laughlin, H. P., *The Ego and Its Defenses,* The National Psychiatric Endowment Fund, eds., second edition (New York: Jason Aaronson, Inc., 1979).
Leary, M., and D. Down, "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, 1995).
Loewenstein, G., and D. Prelec, "Anomalies in Intertemporal Choice: Evidence and Interpretation," *Quarterly Journal of Economics,* CVII (1992), 573–597.
Mele, A., "Real Self-Deception," *Behavioral and Brain Sciences,* XX (1999), 91–136.
Mischel, W., E. B. Ebbesen, and A. R. Zeiss, "Determinants of Selective Memory about the Self," *Journal of Consulting and Clinical Psychology,* XLIV (1976), 92–103.
Mullainathan, S., "A Memory-Based Model of Bounded Rationality," *Quarterly Journal of Economics,* CXVII (2002), 735–774.
Murray, S. L., and J. G. Holmes, "Seeing Virtues in Faults: Negativity and the Transformation of Interpersonal Narratives in Close Relationships," *Journal of Personality and Social Psychology,* XX (1993), 650–663.
Nisbett, R., and T. Wilson, "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review,* LXXXIV (1977), 231–259.
O'Donoghue, T., and M. Rabin, "Doing it Now or Later," *American Economic Review,* LXXXIX (1999), 103–124.
Phelps, E., and R. Pollack, "On Second-Best National Savings and Game-Equilibrium Growth," *Review of Economic Studies,* XXXV (1968), 185–199.
Rabin, M., "Moral Preferences, Moral Rules, and Belief Manipulation," University of California, mimeo, April 1995.
Rhodewalt, F. T., "Self-Presentation and the Phenomenal Self: On the Stability and Malleability of Self-Conceptions," in *Public Self and Private Self,* R. Baumeister, ed. (New York: Springer Verlag, 1986).
Salancik, G., "Commitment and the Control of Organizational Behavior and Belief," in *New Directions in Organizational Behavior,* B. Staw and G. Salancik, eds. (Chicago: St. Clair Press, 1977).
Sartre, J. P., *The Existential Psychoanalysis* (H. E. Barnes, trans.) (New York: Philosophical Library, 1953).
Schacter, D., *Searching for Memory* (New York: Basic Books, 1996).
Seligman, E., *Learned Optimism: How to Change Your Mind and Your Life* (New York: Simon and Schuster, 1990).
Snyder, C., "Collaborative Companions: The Relationship of Self-Deception and Excuse Making," in *Self-Deception and Self-Understanding,* M. Martin, ed. (Lawrence, KS: University Press of Kansas, 1985).
Strotz, R., "Myopia and Inconsistency in Dynamic Utility Maximization," *Review of Economic Studies,* XXII (1956), 165–180.
Swann, W. B., Jr., *Self Traps: The Elusive Quest for Higher Self-Esteem* (New York: W. H. Freeman and Company, 1996).
Taylor, S. E., and J. D. Brown, "Illusion and Well-Being: A Social Psychological Perspective on Mental Health," *Psychological Bulletin,* CIII (1988), 193–210.
Weinberg, B., "A Model of Overconfidence," Ohio State University, mimeo, August 1999.
Weinstein, N., "Unrealistic Optimism about Future Life Events," *Journal of Personality and Psychology,* XXXIX (1980), 806–820.
Zuckerman, M., "Attribution of Success and Failure Revisited, or the Motivational Bias Is Alive and Well in Attribution Theory," *Journal of Personality,* XLVII (1979), 245–287.