Self–Knowledge and Self–Regulation: An Economic Approach [1]

Roland Bénabou [2] and Jean Tirole [3]

Final version, April 2001

[1] We are grateful for helpful comments to George Ainslie, Isabelle Brocas, Danny Kahneman and George Loewenstein.

[2] Princeton University (Department of Economics & Woodrow Wilson School), CEPR, NBER, and IRP.

[3] IDEI and GREMAQ (UMR 5604 CNRS), Toulouse, CERAS (URA 2036 CNRS), Paris, and MIT.

I Introduction

In forecasting how *Homo Economicus* would evolve into *Homo Sapiens,* Thaler (2000) predicted that he would gradually "begin losing IQ", but warned at the same time against simply "making [him] dumber". He recommended instead that economists pay closer attention to the actual processes of human cognition, which psychologists have shown to be subject to a number of specific imperfections and biases.

The research summarized in this chapter can be seen in the light of this overall agenda. Focussing on the links between self–judgement and self–regulation, our approach has been to develop an analytical framework with three main objectives: (i) to unify a number of findings from separate areas of psychology into a parsimonious model of cognition and motivation; (ii) to draw out their main economic implications; (iii) conversely, to derive from the theory potential explanations for the lack of consensus among psychologists on certain empirical and policy–relevant questions, and suggest alternative experimental directions.

II Enriching the psychological makeup of Homo Economicus

We introduce three "grains of sand" (or humanity) into the well–oiled mechanics of the ultra–rational economic agent: *imperfect self–knowledge, imperfect willpower, and imperfect recall.*

Incorporating these three ingredients –sometimes even only a subset of them– yields a surprisingly richer account of human behavior than that of traditional *Homo Economicus.* We are first able to give formal content to individual *traits* such as self–confidence, intrinsic motivation, dependence or autonomy, and power of will, as well as to cognitive *processes* such as wishful thinking or selective memory, self–monitoring, and the setting of personal rules (diets, moral precepts, etc.). The resulting framework then allows us to address the following array of questions. Why do individuals value self–confidence, both for themselves and for others towards whom they are not necessarily altruistic? At the same time, why do they sometimes sabotage their own performance, or deprecate their own accomplishments? Is it possible for a rational, Bayesian individual to deceive himself and hold self–serving beliefs? And if so, is such "positive thinking" ultimately beneficial or harmful? How can we simultaneously account for undersaving and miserliness, procrastination and workaholism, overeating and anorexia? Why do extrinsic rewards (incentives) work well in some contexts, but appear counterproductive in others? Why do agents sometimes undermine the ego and self–confidence of others on whose effort and motivation they depend?

II.1 Imperfect self–knowledge

The agents who usually populate economic models have little doubt about "who they are": they know their own abilities and basic preferences. Information economics is thus primarily concerned with how individuals can *signal to others* (e.g., employers, competitors, tax authorities), these

privately known characteristics, or on the contrary attempt to keep them hidden.[1] Psychology, by contrast, gives a central role to the process of learning about oneself and to individuals' struggle with their own identity: self–esteem, depression, pride, guilt, and self–justification are but a few examples of introspective phenomena. The starting point for our research agenda is therefore the recognition of the fact that people face significant uncertainty about their own abilities and even preferences. That is, they do not know the ultimate costs and payoffs from their actions, and may not even be sure of what actions they would take in a given situation until the very moment when they actually experience it.

Thus, in Bénabou and Tirole (1999a) an individual must decide whether or not to engage (or persevere) in a task which involves sure short–run costs but whose long–run payoff depends on his imperfectly known ability. Typically, the project will be undertaken only if the agent has sufficient self–confidence in his talent, or suitability for the task. In Bénabou and Tirole (2000), the individual is faced with a choice between a course of action which requires no self–restraint (e.g., slack off, drink or smoke as he pleases), and a challenging one where his capacity to resist stress or temptation and hold out for larger, long–run payoffs, will be put to the test (taking on an ambitious project, attempting to quit smoking, going on a diet, etc.). The ability parameter on which uncertainty now bears is the degree to which his preferences may be subject, in certain circumstances, to a bias towards instant gratification.

II.2 Imperfect willpower

Because Economic Man always acts according to his own best interests (given the available information), there is no word in his vocabulary for anything like "self–destructive" behavior, nor any meaning for statements such as "I couldn't help myself". By contrast, a major part of psychology has long been devoted to understanding –and helping alleviate– behaviors characterized by strong internal conflicts, harmful impulses to which the individual succumbs "against his better judgement", or self–deceptions and self–punishments of varying degrees of severity (see Baumeister 2001, this volume). Furthermore, experimental psychologists have documented a robust feature of human time–preferences that commonly gives rise to self–control problems, namely people's tendency to discount payoffs much more steeply at long than at short horizons (hyperbolic–like discounting). In recent years there has been growing recognition by economists of the relevance and explanatory power of such "momentary" preferences which, by creating a conflict of interest between the individual's successive temporal selves, lead to ex–ante suboptimal behavior (prefer-

[1] There are of course some exceptions: see, e.g., Holmström (1982) on workers who learn their abilities over time, Akerlof and Kramton (2000) on the economics of (socially constructed) "identity", or Bodner and Prelec (2001, this volume) on self–diagnostic in a "planner–doer" context.

ence reversal) and create a value of commitment.[2,3]

In Bénabou and Tirole (2000) we expand on the now standard quasi–hyperbolic specification by allowing the degree of present bias (or weakness of will) to be *state–dependent* and *imperfectly known.* Indeed, when deciding whether to go on a diet, embark on an ambitious project (intellectual, entrepreneurial, athletic, etc.) or invest himself in a personal relationship, a key source of uncertainty for the individual is whether he "has it in him" to persevere once the going gets tough. If he is pessimistic as to the likelihood of his eventually caving in to temptation, he will ask himself "what is the point?", and decide that he might as well start indulging himself right away rather than waste effort on a doomed attempt at self–restraint. Thus, once again, the individual's initial self–view is an essential determinant of his behavior. Moreover, because past conduct is a key source of information on one's own willpower, this unknown–preferences feature of the model allows us to capture and better understand the critical role of *self–monitoring* in regulating behavior, as well as the costs and benefits of finding *excuses* for oneself.

One may ask why, if temptation is recurrent, the individual cannot simply and directly recall the preferences (cravings, pain, exhaustion, etc.) which he previously experienced, rather than having to infer them from past behavior and situational factors. The answer relates to the third cognitive "imperfection" from which our simple model of model of Homo Sapiens suffers, compared to his Economicus cousin.

II.3 Imperfect recall and motivated cognition

Standard economic agents may have only limited information but (with a few notable exceptions) what is known or learned is never forgotten nor distorted in their later recollections.[4] In reality, memory is imperfect, attention is limited, and awareness can therefore only be selective. Of particular relevance for our purpose are the following two types of phenomena.

First, a lot of research (not to mention daily observation) has documented the fact that people's recollections of their past actions and performances are often *self–serving:* they tend to remember (be consciously aware of) their successes more than their failures, reframe their actions so as to see themselves as instrumental for good but not bad outcomes, and find ways of absolving themselves by attributing responsibility to others.[5] As explained in Section III.2 below,

[2] See Ainslie (1992) and references therein for the evidence; Strotz (1956), Phelps and Pollack (1968), Laibson (1997) and O'Donoghue and Rabin (1999) for formal models and economic implications. Closest to our approach is the work of Carrillo and Mariotti (2000) and Brocas and Carrillo (1999), which we discuss in Section III.1.

[3] There are alternative ways of representing intra-personal conflict, such a positing multiple contemporaneous selves (Id, Ego and Superego, as in Freudian theory; "Planner and Doer," as in Thaler and Sheffrin (1981) and Bodner and Prelec (2001, this volume); or preferences characterized by the interplay of a "decision utility" and a "temptation utility", as in Gul and Pesendorfer (1999). We adopt (quasi) hyperbolic discounting because of its experimental validation, and because it is the simplest way to parametrize the key concept of *willpower.*

[4] For models with imperfect recall see Piccione and Rubinstein (2000) and Mullainathan (1998).

[5] On self–serving memory, see Korner (1950), Crary (1996), Mischel et al. (1976), Kunda and Sanitioso (1989), or Murray and Holmes (1994). On self–serving attributions, see Zuckerman (1979) and Snyder et al. (1983).

such *motivated cognition* and its compatibility with rational inference represents a main focus of our analysis.

Second, it appears difficult for individuals to accurately recall from "cold" introspection the intensity of stress, temptation or other short–run feelings corresponding to "hot" (visceral, emotional, not easily quantifiable) internal states that they experienced in the past. For instance, Kahneman et al.'s experiments (1997) on the recall of pain and discomfort document a systematic divergence between what they term *experienced utility* and *decision utility.*[6] Such "hot–cold empathy gaps," in the terminology of Loewenstein (1996), also arise in recollections and predictions about feelings such as hunger, exhaustion, drug or alcohol craving, or sexual arousal.[7] Just as people cannot accurately answer retrospective questions such as "how much did it hurt?" or "how much did you dislike that?", we think that asking oneself "how much did I crave that cigarette, or that new dress yesterday?" or "how cold and stiff was I when I went jogging last week?" is unlikely be very informative, compared to asking *what one actually chose to do* –a "revealed preference" approach familiar to economists.

There are of course steps people can take to ensure that they do not forget a piece of data (feelings, actions, circumstances, or outcomes). They can keep written records, rehearse the information, share it with others who will remind them of it, or surround themselves with cues which will trigger the relevant memory. These, however, are *decisions* which the individual may or may not choose to make. Furthermore, the very same strategies can be used to impede or offset the later recollection of unwanted news. There are also ways of interfering with the encoding of such data into memory and their later retrieval, ranging from distracting one's attention to getting drunk "to forget". The natural limitations of attention and decay of memory therefore give the individual some *discretion* about what data is more likely to still be present, later on, on the "computer screen" of conscious awareness.

In line with these ideas and evidence, our model allows for differential rates of recall or accessibility for ego–favorable and ego–unfavorable informations, with the degree of selectivity (relative probability that bad news are forgotten) being either a fixed behavioral parameter, or the endogenous outcome of a (conscious or instinctive) cognitive process over which the individual yields some influence, and in which the benefits of self–esteem enhancement are weighed against the risks of overconfidence. We also allow different types of information to be subject to various degrees to this differential recall. Past internal states (sensations and feelings, i.e. "experienced utility"), being the most "soft" and least verifiable type of information, leads to the most unreliable and manipulable types of memories. External outcomes and circumstances can also be forgotten but less easily so, in the sense that speeding up the forgetting process or building up plausible excuses involves higher costs (destruction of evidence, active cue–management, selective interpersonal in-

[6] The first is measured from the subject's moment–by–moment reports during a painful medical procedure or unpleasant laboratory experiment, while the latter correspond to his or her *retrospective evaluation* of the experience as a whole, and constitutes the informational basis for later decisions. See also Kahneman (2001), this volume.

[7] See Loewenstein and Schkade (1999) for a discussion and survey of the empirical evidence.

teractions, or even substance abuse –drinking to forget). Finally, memories of one's own past actions (especially significant ones) may be the most lasting and reliable, in particular because they tend to leave more material "tracks". These memories also eventually decay, and can be repressed to some extent, but the process is slower and less easily manipulable.

II.4 Rationality

Is our "psychologically enriched" individual, with his imperfect self–knowledge, lack of willpower, and selective recall, still rational? The label in itself is unimportant, but the fact that he retains a rather high level of IQ (cognitive and decision–making sophistication) should be emphasized.[8]

First, we maintain intertemporal utility maximization: at any point in time, the agent tries to do what is best for "himself," given his current (often inaccurate) perceptions of his own interests and abilities. As usual, this optimizing behavior may be interpreted as a good first approximation to what instinct, education or learning will ultimately produce. At the same time, we shall see that even this sophistication does not preclude the agent's finding himself in harmful "self–traps" (inferior personal equilibria).

Second, we do not treat the individual as naive about the systematic incentives to distort his information which he or other people might have. If a principal has a vested interest in an agent's accomplishing a task, the agent will not naively take all encouragements, assurances of success and promises of rewards at face value, but is likely to infer from them some information about the nature of the task or the principal's view of the agent's likelihood of success in the absence of such extrinsic motivators. Similarly, if a person consistently represses or manages not to think about negative news, he or she will likely become aware of this systematic tendency, and realize that the absence of adverse evidence or recollections should not be taken at face value. Formally, this skepticism with respect to others' messages and one's own memories or rationalizations (metacognition) is represented by Bayes' rule. Less sophisticated inference processes lead to similar results, as long as they are not too naive with respect to one's and other people's motives.

In summary, it could be said that in our model each temporal "self" is both optimizing and Bayesian, yet the whole intertemporal collection of such selves which describes the individual is neither, since his behavior is time–inconsistent and his cognition clouded by self–deception.[9]

[8] In a sense, we are still abstracting from a fourth "grain of sand" (much harder to model than the other ones), which corresponds to a limited ability to assess probabilities and make the prospective calculations required to arrive at truly optimal decisions. See, e.g., Gabaix and Laibson (2000).

[9] For psychologists, it may also be worth pointing out that our model is "homunculus–free": at any point in time there is only one set of preferences and information that governs the individual's actions (as opposed to, say a conflicting Ego and Superego, Planner and Doer, etc.).

III Self–Confidence and Individual Motivation

III.1 Why is self–confidence valuable?

In most societies, self–confidence is widely regarded as a valuable individual asset.[10] 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 sizeable part of which purports to help people improve their self–esteem, shed "learned helplessness" and reap the benefits of "learned optimism".[11] American schools place such a strong emphasis on imparting 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?

A first reason may be that thinking of oneself favorably just makes a person happier: self–image is then simply an additional argument in the utility function.[12] Indeed, psychologists emphasize the affective benefits of self–esteem as well as the motivational ones on which we focus, and one might argue that such a "consumption value"of beliefs could have arisen from a more functional (adaptive) mechanism, selected into preferences through evolution or education.[13] To fully operationalize the affective benefits approach, however, would require either an explicit model of preference selection, or empirical evidence imposing structure on the utility–from–beliefs. First, one needs to circumscribe the set of self-images that individuals care about,as there is a potential embarrassment of riches: they may want to perceive themselves as honest and compassionate individuals, good citizens, faithful spouses ... or, on the contrary, pride themselves on being ruthless businessmen, ultra-rational economists, irresistible seducers, etc. Second, one would need to pin down the monotonicity and risk–attitude properties of the utility over beliefs. Whether the individual has a tendency towards blissful optimism or "defensive pessimism" (see Section III.4.2) will hinge on whether this function is increasing or decreasing; whether he is information–averse or information–loving will turn crucially on whether it is concave or convex.[14]

Another limitation of a purely affective theory of self–confidence is that it does not readily extend to *social and economic interactions,* where people clearly seek out optimistic, self–confident partners, rather than depressed, self–doubting ones, and spend substantial time and effort supporting the morale of those they end up matched with. If self–esteem just increases people's utility,

[10] Whether its open display or a more modest, even humble outward attitude is considered socially appropriate is a rather different question, which varies much more across countries.

[11] These last two terms are borrowed from Seligman (1975, 1990).

[12] Following in the path of Akerlof and Dickens' (1982) celebrated model of dissonnance reduction, a number of recent papers such as Rabin (1995), Weinberg (1999) and Köszegi (1999) have also introduced self–beliefs as arguments of individual preferences.

[13] For instance, the overconfidence that often results from such preferences may propel an individual to undertake activities (exploration, foraging, combat) which are more risky than warranted by their private material returns, but nonetheless evolutionarily successful because they confer important external benefits on the species.

[14] See Caplin and Leahy (1999) for a study of attitudes towards information and the resolution of uncertainty for a general class of preferences where beliefs over future lotteries enter into the intertemporal utility function.

something else must account for why we like not just ourselves and our family members to be self–confident, but also our coworkers, managers, employees, teammates, soldiers, and numerous others with whom we have only functional, non–altruistic interactions.

By contrast, the motivation–based theory which we develop offers a unified explanation of why (and when) self–confidence is valuable for ourselves *and* for others with whom we interact. The basic idea is that since a more optimistic view of his own abilities generally enhances an individual's expected return from effort, anyone with a vested interest in his performance has an incentive to build up and protect this self–esteem. This occurs in two classes of situations, corresponding respectively to *externalities* and *"internalities"* from the agent's effort.[15] First, the manipulator could be another person (parent, teacher, spouse, friend, colleague, boss or relation) who would benefit from the agent's performance; such *interpersonal* issues are studied in Bénabou and Tirole (1999b), and will be briefly discussed in Section V. Second, as pointed out in Carrillo and Mariotti's (2000) seminal paper, an individual suffering from time inconsistency often has incentives to restrict his own information, as a way of indirectly controlling the behavior of his future selves. Building on this idea, we model in Bénabou and Tirole (1999a) the following "canonical" problem of *self–motivation.*[16]

An individual would like, ex ante, to pursue a certain desirable project (professional, intellectual, athletic, health–related, etc.). However, he realizes that, due to his bias towards immediate gratification, he is too likely to give up along the way, thereby sacrificing significant expected long–run benefits for short–run relief from the effort costs of completing the task. That is, while he may not know in advance the effort cost that his future "self" will face, he knows that the range of cost realizations for which the task will be abandoned is suboptimally large from a long–run point of view. At the same time, his sustained motivation for such endeavors will naturally depend on his assessment of his chances of success, or more generally of the costs and benefits he is likely to incur or reap from their pursuit. As a result, the individual's current "self" has a vested interest in maintaining and enhancing the self–confidence of future selves, so as to counter their natural tendency to procrastinate or give up too easily. Imperfect self–knowledge and imperfect willpower thus combine to generate an endogenous, instrumental value of self–confidence. We discuss below a first round of implications involving a priori attitudes towards information, then turn to the issue of self–deception (ex–post belief manipulation) and present the main results obtained from the interaction of these two building blocks of our theory.

1) Receptivity to information. Because it counterbalances the deleterious effect of present–orientation on his motivation for effort, self–confidence is a valuable asset for the individual. Consequently, he may prefer to remain blissfully ignorant of his true abilities and past achievements than to endanger it, even when accurate information is freely available. This is of course a direct application of the general "strategic ignorance" principle for time–inconsistent agents

[15] The term "internalities" is due to Herrnstein et al. (1993).
[16] See also Brocas and Carrillo (1999; 2001, this volume).

in Carrillo and Mariotti (2000). Conversely, someone with very low initial self esteem will be desperate for good news that might lift him out of the "procrastination trap", and his choices of tasks and social interactions will have the nature of "*gambles for resurrection*" of his self–esteem.

2) Self–handicapping. It is not uncommon for people to sabotage their own performance –a behavior which, on its face, appears inconsistent with the rational pursuit of self–interest.[17] In experiments, subjects with high but fragile self–confidence choose to take performance–impairing drugs before an intelligence test. In daily life, people withhold effort, prepare themselves inadequately (e.g., getting drunk or not sleeping enough the night before an exam), or select overambitious tasks where they are almost sure to fail. Test or performance anxiety is another example of self–handicapping behavior. Psychologists such as Berglas and Baumeister (1993)) have suggested that self–handicapping is often a self–esteem maintenance strategy (instinctive or deliberate), directed both at oneself and at others. For economists, our model provides a validation of this insight by showing that it can be quite rational to sacrifice current performance in order to reduce the probability of a large negative inference about one's self. For psychologists, our analysis has implications for experiments on self–esteem maintenance, to which we now turn.

3) Implications for experimental research. There has been a significant amount of work examining whether it is high– or low self–confidence individuals who are most likely to engage in self–esteem maintenance strategies such as deliberate ignorance and self–handicapping. Although there seems to be somewhat more evidence in favor of the first hypothesis, our reading is that the sum of these experiments has not yielded any firm conclusion. Our model shows that this ambiguity is in fact to be expected, and suggests instead another variable whose correlation with such behaviors should be tested.

First, subjects' point–estimates of their abilities, whether self–reported or measured by scores on a self–esteem scale, are not sufficient statistics to predict attitudes towards information. Basic intuition makes clear that the extent of uncertainty around this estimate plays a crucial role; a similar point was already noted by Greenier et al. (1995). Furthermore, the analysis reveals that it is not even the variance of one's priors that matters, but a more subtle feature of the distribution (a likelihood ratio property). Such detailed information about how uncertain a person is of his own traits would not be easy to measure experimentally; and even if one could do it, there would remain a major difficulty: the value of information (the amount an agent is willing to sacrifice to obtain or avoid it) is typically not monotonic in his initial self–confidence.[18] Intuitively, those with very negative self–assessments have nothing to lose from information, while those with very positive and secure self–views have nothing to fear from it. For those with moderate and diffuse self–views, on the other hand, information may be dangerous.

[17] See, Berglas and Jones (1978), Arkin and Baumgardner (1985), Fingarette (1985) or Gilovich (1991).

[18] The one case where a firm conclusion may be drawn is when information is available, or avoidable, at zero or minimal cost. Individuals with higher self–esteem (in a monotone likelihood ratio sense) are then always the more likely ones to accept a signal about their ability, because they have more at stake in terms of motivation.

Rather than initial self–esteem, the personal trait which the model suggests should be robustly correlated with a subject's propensity to engage in willful ignorance, self–handicapping and the like is their degree of degree of undermotivation, that is, their (self–perceived) bias towards instant gratification. The economic approach to self–regulation –common to our work and that of Carrillo and Mariotti (2000) and Brocas and Carrillo (2001)– thus suggests a link between two hitherto disjoint areas of experimental psychology, namely those on intertemporal preferences and self–esteem maintenance.

III.2 Can rational individuals deceive themselves?

Like it or not, in daily life we are subjected to a constant flux of feedback about our performance and abilities, be it from parents, teachers, spouses, coworkers, etc., or simply from observing our own performance and comparing it to that of others. The relevant issue is thus often not whether to seek or avoid information ex–ante (i.e., before knowing what it will turn out to say), but how to cope with the good and especially the bad news that one inevitably receives.

The answer might appear obvious: as Mark Twain succinctly put it, *"Denial aint' just a river in Egypt."* Indeed psychologists, and before them writers and philosophers, have long documented people's universal tendency to deny, explain away, and selectively forget ego–threatening informations. Freudian repression into the unconscious is the now unfashionable archetype (or perhaps caricature) of such behavior, but various other forms of *motivated cognition* and *self–deception* feature prominently in contemporary psychology.[19]

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

Our dynamic model allows us to unbundle the "self that knows" from the "self that doesn't know", and thereby reconcile the motivation and cognition aspects of self–deception within a standard information–theoretic framework. The basic idea is that the individual can, within limits and at a cost, *affect the probability of remembering* a given piece of information. Under time–inconsistency, there is an incentive to try and remember signals that help sustain his long–run goals, and to forget those that undermine them. This is the motivation part.[20] 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.

To make things more concrete, suppose that the individual wishes to remember good news and forget bad ones. He can linger over praise or positive feedback, rehearse them, and choose to

[19] See the references cited in Section II.3. For a general discussion of self deception, see Baumeister (1998).

[20] Alternatively, it could arise from a purely affective value of self–esteem. The awareness–management component of our model could thus also be combined with a utility–for–beliefs approach.

be more frequently in environments or with people who will remind him of his past successes.[21] Conversely, he can eschew situations and people who remind him of his failures, tear up the picture of a former girlfriend, or work unusually hard to "forget" (really, not think about) a failed relationship or family problem –even use drugs and alcohol. The individual can also use a wide array of strategies to discount the bad news in the first place (question the motives of the news–bearer, search for contradicting evidence) or impair their accurate encoding and recollection (e.g., create a distraction, such as an emotional outburst which moves the interchange from "talking to each other" to "talking past each other").[22]

It is thus important to note at the outset that we need not *literally* assume that the individual can directly and mechanically suppress memories. Our model is equally consistent with a Freudian view where memories get buried in the unconscious (with some probability of reappearance), and with the more recent cognitive psychology view which holds that memory itself cannot be controlled, but emphasizes the different ways in which *awareness* can be affected: the choice of attention when the information accrues, the search for or avoidance of cues and the process of selective rehearsal afterwards, and again the choice of attention at the time the information is (voluntarily or accidentally) retrieved. While the means employed may be very different, the end–result of these two views of motivated belief formation is formally equivalent: the individual has differential rates of recall, or accessibility, depending on how helpful or hurtful the information is to his self–esteem and general efficacy.[23]

III.3 Intrapersonal equilibrium

Recall the basic predicament of our time–inconsistent individual, whose bias towards immediate gratification means that he will always be too tempted (in a long–run welfare sense) to give up, procrastinate or altogether shy away from hard tasks with delayed payoffs. As explained earlier, a greater degree of confidence in his likelihood of success if he does persevere, or in the size of the prize that this will yield, will tend to alleviate this underprovision of effort. Conversely, bad news about the expected payoffs from the project will further undermine his motivation.

Suppose the individual has in fact just received such bad news. In deciding, consciously or instinctively, whether to try and censor them, he faces a basic tradeoff between preserving the effort motivation of tomorrow's self (gain from confidence–maintenance) and the risk of becoming overconfident, meaning that he will blindly persevere even in circumstances so adverse that it would be (ex–ante) optimal to withdraw. Within the limitations of the available "technology"

[21] See Rhodewalt (1986) for a discussion of such self-presentation strategies and their link with self- enhancement.

[22] As Gilbert and Cooper (1985) argue, "social interaction is a fertile context for self-deception because its very complexity often acts as a "smoke screen", keeping the self-deceptive process from becoming obvious."

[23] It is also worth pointing out that allowing for the *possibility* of selective awareness or memory in the model's assumptions does not prejudge in any way whether (or when) such wishful thinking will actually be used. Our endogenous–memory, Bayesian model is thus very different from one where agents have a fixed, mechanical tendency to optimistically bias their interpretation of all self–relevant signals.

of awareness manipulation, and given the costs involved, the individual's current self chooses the probability of recall so as to optimally balance these two effects. Consider next the inference problem faced by tomorrow's self: given that bad news are forgotten with a higher probability than good news, what credence should be given to the fact that only good news can be recalled? Assuming that the individual is not completely naive about his own motives and habits, his self–confidence and decisions will take into account the imperfect reliability of his memories. Conversely, the degree of selectivity chosen at any point in time will depend on how "skeptical" the individual expects to be in the future. These two problems are interdependent, leading to a *game with private information* between today's and tomorrow's self.

We first highlight some important general features of the set of behaviors that will typically result, then turn in Sections III.4 and III.5 to the more specific questions of optimism, pessimism, and the welfare consequences of self–esteem maintenance.

1) Self–traps. A first result emerging from the analysis is that multiple modes of cognitive behavior are often self–sustaining for the same person, or for otherwise identical subjects. Thus, an intrapersonal equilibrium with a low level of repression (self–honesty) often coexists with another one characterized by a high level of repression (positive thinking, denial), and an intermediate case in–between. Unless the individual can successfully coordinate the strategies and expectations of his temporal selves (a point on which we are agnostic), he or she may well be trapped in an inferior equilibrium.[24,25]

2) Self–doubt: repressing bad news more systematically makes recalling only good news more suspect, ultimately impairing the credible transmission of both types of signals. Just 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.[26]

3) Testable implication. The model also shows that the tendency to engage in selective memory and similar forms of self–deception is greater for more time inconsistent (weak–willed) individuals. Indeed, the benefits of confidence–building rise with time inconsistency, while the risks of overconfidence decrease with it.

[24] What makes all three equilibria self-fulfilling is thus precisely the *introspection* or "metacognition" of the Bayesian individual, who understands that his cognitive process (in this instance, his memory) is subject to opportunistic distortions. The higher the degree of censoring by today's self, the more tomorrow's self 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 today's self incentive to censor. Conversely, if today's self faithfully records all news in memory, tomorrow's self is more likely to be overconfident when he cannot recall any bad signals, and this incites today's self to be truthful.

[25] A closely related result is that small causes can have large effects: minor variations in preferences or incentives (time discounting, psychic or material costs of memory management, repression, etc.), can lead to large and sudden changes in cognitive strategies, self–confidence, and ultimately in performance.

[26] This coarsening of information is quite different from the a priori suppression of signals seen earlier with self-handicapping, or in Carrillo and Mariotti (2000). In particular, we shall see that it may end up doing the individual more harm than good, whereas the usual "strategic ignorance" is only chosen when it improves ex-ante welfare.

III.4 Optimism and pessimism

III.4.1 The prevalence of self–serving beliefs

Surveys, experiments and common observation consistently suggest that most people overestimate their past achievements, abilities and other desirable characteristics, both in absolute terms and relative to others (e.g., Weinstein (1980), Taylor and Brown 1988)). While this is often taken as evidence of pervasive irrationality in human inference, it turns out that *rational self–deception* by Bayesian agents can very well account for most people holding biased, self–serving beliefs.

To take a concrete example based on our endogenous–memory model, suppose that 2/3 of the population receive a signal that they are of low ability, while the remaining 1/3 receive no such signal (no news is then good news). If the costs of repression and awareness management are low enough, a large fraction, say 3/4, of the low–ability group may successfully repress these bad news. Because of the self–doubt effect they will not be as optimistic as if they were certain that no adverse signal had occurred, but nonetheless the suppression of the adverse data will raise their self–assessments above the unconditional prior, which is also the population average. As a result, a majority of the total population ($3/4 \times 2/3$) will believe themselves to be *more able than they actually are, more able than average, and more able than 2/3 of individuals.* Adding the 1/3 who had truly received experienced good news (not received a negative signal), the fraction who think they are better than average is even larger, namely 5/6; note, in passing, that the 1/3 able agents actually underestimate their true talent, due to the self–doubt effect. The remaining 1/6 of the population think, 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 to their non–depressed counterparts who are much more likely to exhibit self–serving delusions (Alloy and Abrahamson 1979).

The key intuition here is that Bayes' law does not constrain the skewness in the distribution of biases.[27] It only requires that the *average bias* across the overconfident and underconfident agents in the population (who number 1/2 and 1/3 in our example) sum to zero. Note also that, since decisions are typically non–linear in beliefs, a zero average bias in no way restricts self–esteem maintenance strategies from having aggregate economic and welfare effects.

III.4.2 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 studying 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 to what will be required to obtain

[27] This was first pointed out by Carrillo and Mariotti (2000) in a context of ex–ante strategic ignorance. See also Brocas and Carrillo (1999; 2001, this volume) and Köszegi (1999).

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 *defensive pessimism* can be captured with a very simple variant of our basic 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 will typically occur when the reward for performance is of a "pass–fail" nature, such as in obtaining a diploma, making a sale, being hired or fired (tenure, partnership) –perhaps also in marriage and divorce. Note that this yields another *testable* prediction of the model, from which it could be distinguished from a purely hedonic theory of self–confidence: one could for instance compare subjects' confidence–maintenance behavior across experiments (or careers) where payoffs were complements and substitutes.

III.5 Is "positive thinking" good for you?

As noted earlier, there is an enormous industry of "self–help" books, courses, gurus and now web sites claiming to help people improve their self–esteem and that of their children. But is a person ultimately better off following a strategy of active self–esteem maintenance and "positive thinking," or when he always faces the truth? Psychologists –both researchers and practitioners– appear sharply divided between these two conflicting views of self–deception. On one side are those who endorse and actively promote the self–efficacy / self–esteem movement (e.g., Bandura 1977, Seligman 1990), pointing to studies which tend to show that a moderate dose of "positive illusions" has significant affective and functional benefits. On the other side are skeptics and outright critics (e.g., Baumeister 1998, Swann 1996), who see instead a lack of convincing evidence, and point to the dangers of overconfidence as well as the loss of standards which results when negative feedback is systematically withheld or discounted in the name of self–esteem preservation. Our formal analysis helps provide insights into the reasons for this ambiguity.

Recall that the benefit of a "hear no evil–see no evil" strategy is that it helps the individual preserve motivation in the event of bad news. It does involve a risk that one will be overconfident, but the individual is aware of this tradeoff, and only censors signals when it leads to expected utility gains. This net *gain from forgetting bad news* is only one side of the coin, however: the other one is the *loss from disbelieving good news,* due to the self–doubt effect explained earlier. Thus, while the individual will be better or even over–motivated following a negative signal about his ability, he may actually be undermotivated following a good signal. Which effect dominates ex–ante welfare therefore depends on the general difficulty of the task (the distribution of effort costs) and on his degree of time–inconsistency. Our analysis establishes that:

1) When the tasks one faces are very difficult relative to one's willpower, an active strategy of self–esteem maintenance, selective memory, "looking on the bright side", etc., can indeed pay off.

2) When the typical task is likely to be only moderately challenging, and time inconsistency is not too high, one can only lose by playing such games with oneself , and it would be better to always "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.

IV Personal rules and tests of willpower

IV.1 Reasoned self–control

The cognitive strategies discussed until now represent attempts by the individual to manipulate his future self's *perception* of the payoffs attached to alternative courses of action: work or play, persevere or give up, abstain or drink, etc. Through strategic ignorance, self–handicapping, selective recall and the like, he aims to minimize the intensity of future temptation, that is, to reduce the divergence between his ex–ante (or long–run) and his ex–post (or temporary) preferences. A closely related mode of self–regulation is *preparation of emotion*,[28] where the agent trains himself to care less about certain desires, or even to find them repulsive, by associating them with vivid, disgusting images (e.g., cigarettes and fatty foods with visions of diseased lungs and clogged arteries), and conversely by pairing positive images with delayed–gratification actions (receiving an award, achieving fame, etc.).

While undeniably important, these behaviors still fall short of fully capturing what is usually meant by self–control, namely the deliberate, reasoned and contemporaneous *overriding* of impulses, at the time they occur.[29] Thus, Baumeister et al. (1994) observe that most forms of self–control involve interrupting or suspending a naturally occurring physical or mental response to a desire or craving, and conclude that *"the essential nature of self–regulation is that of overriding."* Similarly, Ainslie (1992) contrasts the use of either external commitments, which takes away the tempting option, or emotion control, which reduces its attractiveness, to *"the kind of impulse control which we call willpower, which allows a person to resist impulses while he is both attracted by them and able to pursue them".*

So how do people resist the temptation which they are *currently* experiencing, and can our simple economic model also help understand these more direct forms of self–regulation? The answer likely begins with the typical rationalization which accompanies *failed* attempts at self–restraint: "just this time..." Conversely, reasoned self–control requires perceiving a clear link between behavior today and behavior in the future, which transforms the impulsive act one is

[28] This term is used by Ainslie (1992), and clearly emphasizes the ex ante nature of this strategy.

[29] By contrast, in what precedes the current self never attempts to restrain his own impulses, but only invests in information or disinformation that will help counteract the anticipated impulses of future selves.

about to commit from an isolated decision into a *precedent* which bodes ill for all future such choices. "If I eat this tempting desert, there goes my whole diet. If I cannot turn down this drink, I might as well admit that I am still a hopeless alcoholic." As illuminatingly explained by Ainslie (1992, 2001), the way out of the time–inconsistency trap is to view choices between rewards not one by one, but as indissociable parts of more permanent patterns of behavior (impulsive versus reasoned, weak versus strong, moral versus immoral etc.), which either affirm or violate long–lasting rules to which one would like to conform. Indeed, the setting and monitoring of *personal rules* and targets for oneself is perhaps the most prevalent strategy used by people to try and achieve self–restraint. Examples include diets, resolutions to smoke only after meals, jog four times a week, write five pages a day, always finish what you started, conduct your life in a way that would have made your mother proud, and all sorts of similar "promises to oneself." The question is, of course: given that these rules are entirely self–imposed, how (or when) can they actually constrain the individual's behavior? And, in particular, *why* should misbehavior today make it more likely that one will also misbehave tomorrow?

IV.2 Self–monitoring: inferring one's preferences from one's actions

We have not found in our reading of the psychology literature a fully spelled out answer to these questions. Ainslie (1992) provides the most explicit clues, which suggest an important role for uncertainty and learning about one's preferences: "*In situations where temporary preferences are likely, [the individual] is apt to be genuinely ignorant of what his future choices will be. His best information is his knowledge of his past behavior under similar circumstances...Furthermore, if he has chosen the poorer reward often enough that he knows self–control will be an issue, but not so often as to give up hope that he may choose the richer rewards, his current choice is likely to be what will swing his expectation of future rewards one way or the other.*" Together with the emphasis put by other psychologists such as Baumeister on the importance of self–monitoring for successful self–regulation, and on how devastating to a subject's self–view and subsequent behavior breaking a strict personal rule can be ("lapse–activated snowballing"), this quite naturally leads us to propose a theory of *self–reputation* over one's willpower, defined as the inverse of the bias towards immediate gratification.

The basic idea is that by breaking the rule the individual would *reveal himself,* in his own (future selves') eyes, as weak–willed –that is, incapable of resisting temptation. Such a loss in self–reputation would further undermine his resolve in the future, to the point where he may even abandon all attempts at self–restraint: what is the point of sticking to a diet today if, based on recent experience, it is likely to be broken tomorrow? Because of learning there can be no "just this once," and the threat of triggering a significant and lasting loss of self–control can help limit or even override the individual's natural bias towards immediate gratification. While this intuitive description is still incomplete (we shall develop it further below), it already makes clear two important points.

First, a personal rule's power is predicated on the likelihood that a lapse today will later on be recalled and interpreted as denoting weakness of will, rather than forgotten or rationalized away (finding an excuse). Thus, the accuracy and reliability of the individual's monitoring and interpretation of his own actions are essential. Indeed, psychologists observe that most failures of self–regulation such as overeating, addiction, procrastination, etc., are associated with inaccurate or deficient self–monitoring (Baumeister et al. 1994). Conversely, simply adopting or being forced to adopt a more regular and accurate self–monitoring "technology," such as keeping a journal of one's successes and lapses, often leads to a reduction in the occurrence of impulsive behavior (Ainslie 1992). This will indeed be an implication of the model.

Second, and related to the first point, the fact that people commonly monitor and draw inferences from their own actions reveals that they must suffer from a deeper form of imperfect self–knowledge than the one discussed in previous sections. Whereas earlier they knew the underlying motives for their behavior but not all of its future consequences, they must now be uncertain about their very own preferences. Indeed, any time a person looks back to his past actions to infer what he is likely to do in the future, it *must* be that the preference ordering (motive) which led to the earlier decisions has some permanence (making it relevant to future choices), but nonetheless can *no longer be recalled* or accessed with complete accuracy or reliability. This kind of inaccessibility of one's preferences is quite different from the more transient kind of taste uncertainty found in economics, such as not knowing whether one will like a new food, product or job (experience goods, learning by doing).[30] It is, on the other hand, very much in line with the evidence discussed earlier about people's inability to recollect their own past utilities, pain, cravings etc. More generally, most psychologists view introspection as a very imperfect source of self–knowledge, leading individuals to commonly infer or reconstruct their own motives from their past actions (retrospective justification).

IV.3 Personal rules and self–reputation over willpower

Bringing together this convergent set of findings and ideas from psychology, we model in Bénabou and Tirole (2000) the behavior of individuals who are unsure of their willpower (ability to delay gratification) when under stress or confronted with intense temptation, and must therefore infer it from the extent to which their previous conduct conformed to, or strayed from, more or less stringent rules. We characterize the rules which can be sustained as intrapersonal equilibria where impulses for immediate gratification are held in check by the fear of "losing faith in oneself" (damaging one's self–reputation), which would lead to a further collapse of self–discipline. We also examine how they depend on the extent to which the individual's self–monitoring is subject to opportunistic distortions of memory or inference (such as finding excuses for oneself) of the type discussed in Section III.2.

[30] For instance in Carrillo (1998), a time–inconsistent individual may choose total abstinence with respect to goods such as alcohol or drugs in order to never find out that he (perhaps) finds them too tempting to resist.

The basic model of imperfect self–knowledge, willpower and recall outlined in Section III.3 is thus extended as follows. The two basic periods –present and future or, for concreteness, "today" and "tomorrow"– are each subdivided into two subperiods –say morning and afternoon. Each morning, the individual first decides whether or not to try and exercise self–restraint that day. Trying means embarking on a willpower–dependent activity, where his capacity to defer gratification (or withstand discomfort) will be put to the test later in the day (second subperiod). Not trying means indulging his impulses right from the start, or more generally choosing some activity where willpower will not be tested. Thus for a smoker, attempting to exercise willpower means first abstaining from lighting up in the morning, and then trying to hold fast through the afternoon, when the craving may become much more intense, or the temptation be heightened by the proximity who smoke. For a procrastinator, it means first setting to work on his book in the morning, and then trying to keep at it in the afternoon, when fatigue, boredom and possible distractions will reach their peak. While the first phase of self–restraint is relatively easy, the second one is typically much harder, and the individual is generally uncertain of whether or not he will have the willpower to ride it out.[31] Furthermore, if he expects to give up later in the day, the delayed benefits (incremental health, pages written) from only one morning of "good behavior" may well be insufficient to compensate for his initial proclivity towards immediate gratification; in that case he will not even make the initial effort, but indulge his impulses right from the start. This, in turn, implies that much more than a half day's worth of delayed rewards rides on his persevering during the first afternoon. The individual knows that if he caves in (and if no extenuating circumstances can be invoked), he risks *appearing as weak–willed to himself* the next day, and his demoralized future self will then abandon all restraint and just start smoking or shirking first thing in the morning. A lapse today thus indeed acts, through its informational spillover, as a *precedent* which leads to a further deterioration in behavior tomorrow. Aware of these high stakes, the individual will then exercise greater self–restraint in the face of temptation. Thus are rules such as "I don't smoke anymore" or "I work on my book ten hours a day" sustained through the force of self–reputation.[32]

Our simple model thus allows us to identify the key ingredients and mechanisms for a psychologically grounded, mathematically explicit, theory of rule–based behavior. In what follows, we emphasize some of its main results.

[31] See Baumeister (e.g., 2001, this volume) for evidence that willpower –which he compares to a muscle– depletes over time with stress and fatigue.

[32] The two–period setup is of course only a simplifying abstraction. A longer horizon would only amplify the effects identified here, as each lapse would put many periods' self–restraint at risk. Note that our non–cooperative, reputation–based model could be seen as providing an informational foundation for Caillaud et al.'s (1999) cooperative game–theoretic approach to self–restraint (cooperation or noncooperation here is between temporal selves).

IV.4 Lapses as precedents

1) The value of self–confidence once again. The model shows that the extent of self–discipline an individual will achieve (keeping constant his true preferences) is generally higher, the greater his confidence in his own willpower.[33] The intuition is that a greater reputational capital is then at stake in each decision, resulting in more forward–looking behavior. As a result, ex–ante welfare is also higher. Thus, as with the assessment of one's productive ability, there arises an endogenous, instrumental value of confidence in one's "character".

2) Dependence. A closely related result concerns the notion of dependence and the conflict between extrinsic and intrinsic motivation.[34] Suppose that, during an initial phase (e.g., childhood), the individual's behavior is subject to tight external constraints –imposed for instance by "controlling" parents or a society with rigid norms. He will of course behave better, but be deprived of any opportunity to test his will and build up a reputation sufficient to ensure successful *self*–regulation later on, when he is left to his own devices. The externally enforced initial "good behavior" thus comes at the expense of later autonomy (ability to make the right choices by oneself) and self–restraint. An excessively "protected" childhood has the same effects.

3) The role of self–monitoring. Once a lapse (succumbing to temptation) *has* occurred, it is a piece of "bad news" about himself which the individual will often have strong incentives to try and forget or "explain away", so as to avoid the costs associated with a damaged self–reputation. To capture this well–documented self–serving bias in memory we allow, much as in Section III.2, each lapse to be forgotten with positive probability. We show that only when the recall probability is sufficiently high (in a well–defined sense) will behavioral rules with actual self–control be sustainable. There arises therefore a strong tension between the individual's ex–post incentives to forget lapses, and his ex–ante knowledge that he can only escape "being the slave of passion" if he is able to prevent or limit such selective memory or attention. This points to the importance of choosing rules with good mnemonic properties (e.g., only one cigarette after each meal, rather than just three per day) and of rehearsing them often (e.g., religious principles) so as to *make lapses more salient* if they do occur. As noted earlier, people who fail at self–regulation are often those with poor self–monitoring, and conversely much of behavioral therapy involves endowing the patient with a more reliable, less manipulable "technology" for monitoring his actions, such as keeping a journal or talking regularly with the therapist. Twelve–step programs and weekly confession in church are also, in large part, self–monitoring devices. Neither his peers in the group nor the priest in the confessional can (or even try) to force the alcoholic or sinner to refrain from his vice (contrast this to, say, a rehabilitation clinic). Admittedly there is some social or moral pressure, but the individual could avoid it by lying about his recent drinking or bad actions. He may perceive some probability of being found out and condemned by his peers,

[33] As long as the problem of "false excuses," which we discuss in the next section, is not too severe.

[34] See Section V.1 below for psychological references and further results in an interpersonal context.

or punished by divine intervention, but the direct and inevitable effect of participating in such "pure talk" sessions is that the individual is *forced to think about his own behavior,* even if he chooses to misrepresent it.

IV.5 Excuses, exceptions and compulsiveness

In sufficiently adverse circumstances, even the strongest–willed individual would, and perhaps even should, give up rather than persevere. When really feeling sick, or learning that a friend in trouble needs comfort, even a perfectly time–consistent person will postpone work to another day. If the weather is excessively cold, the jogging–every–day rule should be broken; if the host insists that one have some of his or her special desert, it would be more impolite than heroic to refuse. In inferring one's strength of will from one's actions, attention must thus be paid to the circumstances under which they took place. This *signal–extraction* problem is further compounded by the fact that, once again, memory may be self–serving: an individual who recalls straying from his rule will generally have a strong incentive to come up with plausible "excuses" which allow him to attribute the lapse to temporary, external factors, rather than enduring, personal ones.

These issues can be examined in our model by allowing the cost of perseverance (craving) to be stochastic: on any given day (afternoon) it can be either high, which constitutes a valid "excuse" for giving in, or low, in which case there is no excuse, meaning that only a weak–willed individual would ever give up. Even in the latter case, however, the individual may still manage to come up ex–post (with some probability) with some plausible justification. Not being completely naive he will realize that this recollection could be self–serving rather than genuine, but this ambiguity is still better than not having any admissible excuse at all.

1) Patterns of self–regulation. The variability of situational factors allows us to distinguish between *flexible rules* (persevere except in really adverse circumstances) and *rigid rules* (never give in – take no excuses). The former corresponds to what an individual with no commitment problem would generally do. It can also be adopted by one with relatively weak willpower who, by imitating ("pooling with") the strong–willpower type, is able to better restrain his own impulses. More novel at this point is the rigid type of behavior which occurs when an individual who has relatively high willpower, but is insecure about it, feels compelled to "prove himself" in every instance by tolerating no exception to his rule.

2) Compulsiveness. The type of stringent rule described above could sometimes (depending on parameters), be desirable a priori. More interesting is the case where perseverance in the high–cost state is undesirable even ex–ante, that is, would never be chosen by a fully time–consistent, self–knowledgeable individual. In that case it represents an unambiguous cost which the individual incurs for the sole purpose of reassuring himself about his own character. Such excessively rigid or "legalistic" rules correspond well to *compulsive* or *obsessive* behaviors such as those of the miser, the workaholic or the anorexic: the individual is so afraid of appearing weak to himself that every decision becomes a test of his willpower, even when the stakes are minor or when self–restraint is

actually harmful.

3) Overregulation and underregulation: two sides of the same coin? The preceding result is important because it shows that the model is indeed able to account for over–as well as under–regulation within the same, simple, framework. Hyperbolic–type preferences are sometimes criticized for failing to explain common instances where individuals seem to *over*weigh distant payoffs; we show that this apparent "salience of the future" is not only consistent, but actually generated by (a concern over) present–oriented preferences. Ainslie (1992) conjectured that compulsiveness is a "side effect" of personal rules edicted to alleviate weakness of will. Baumeister et al. (1994, 85–86) summarize a similar view held by many psychologists, namely that *"Obsessions and compulsions are attempts to compensate for some self–regulatory deficit... The quest for such structure [boundaries, limits, time markers, and the like] and the excessive adherence to such structure, which have been commonly observed among these individuals, may be a response to the inner sense that they cannot control themselves without those externals aids."* Our model spells out formally the common cognitive mechanism through which *both* deficient and excessive self–restraint may occur. It also yields testable predictions, such as the fact that compulsive behavior is more likely when the individual's initial self–reputation is low relative to his true willpower, and when the veracity of self–excuses and ex–post rationalizations is difficult to ascertain (because he then does not "trust his own judgement").

4) Credibility of excuses. Indeed, suppose that the environment in which the individual operates becomes more permissive of (or conducive to) self–serving retrospective justifications: say, it is harder to ascertain later on whether extenuating circumstances did or did not apply; or, there is less probing and skeptical questioning of the individuals's rationalizations by others (e.g., an experimenter). In terms of our model, this corresponds to a higher probability that a plausible (non–falsifiable) excuse can be found even when no extenuating circumstances existed for a previous lapse in behavior. How will this affect the extent of self–restraint by the individual? Our analysis suggests a clear dichotomy:

– people with relatively low actual willpower (high susceptibility to temptation) respond by "abusing" the excuses, and therefore *lose* self–restraint. Furthermore, this loss of is more drastic, the *higher* the initial self–confidence. The intuition is that when the self–monitoring "technology" is not very effective, a weak type with a good self–reputation can take full advantage of his "principal's" (tomorrow's self) high level of trust, and misbehave without any adverse consequences. When initial reputation is low, however, tomorrow's self will apply a tougher standard, so the weak type must actually "work at" passing for a strong one (exercising self–restraint).

– for those with relatively high actual willpower, by contrast, less reliable inference leads to a *tightening* of self–restraint, or *increased compulsiveness.* The idea is that the compulsive individual is trying to prove himself by systematically doing things "the hard way"–that is, by taking actions that would be too costly for a weak–willed individual to mimic. When false excuses become easier to come by, he can only maintain a tough standard for himself by refusing to accept

exceptions to the rule altogether.

These implications of the model should again be testable, provided one can: a) assess subjects' initial self–perceptions of their willpower;[35] b) vary the extent to which the environment in which they must resist temptation offers arguments that could be invoked later on as excuses (e.g., tell them that most people give in, or do not give in; that this is a hard, or an easy task; visibly monitor, or not monitor, the conditions under which the test is done).

V Social interactions

A unified approach to social psychology should start from a single view of the individual's preferences, cognitive machinery and basic problem–solving strategies. While incentives and feedback, and therefore behavior, are highly context–dependent, the underlying "fundamentals" are the same whether the individual is engaged in self–regulation or interacting with others. Although a single unified model is surely an unattainable ideal, we briefly point out in this section how our simple theory, based on the interplay of self–knowledge and motivation, can also shed light on a number of psychological findings related to social interactions.

The common thread running through a wide variety of social situations is that one agent (or more) is trying to get another one (or more) to perform a certain task: study, work, buy or sell, consent to a relationship, etc. Conversely, the other party is interested in determining, and if possible maximizing, "what is in it" for them. In such settings, which economists refer to as principal–agent relationships, psychologists have studied two types of interactions (going in opposite directions) between an individual's self–view and his social environment.

1) *Self–presentation.* Here, the individual attempts to manipulate others' behavior by conveying information about himself through self–promotion, intimidation, ingratiation, excuse–making, supplication, and the like. What allows such signalling is that an individual typically has private knowledge relevant to assessing his abilities, such as memories of his past performances and of how they were achieved (intensity of effort, circumstances that affected the results). For concreteness, suppose that this individual, called the "agent" (he), interacts with a "principal" (she), who can alter the attractiveness of a task (rewards, punishment, difficulty) for the agent, or even decide on the set of allowable tasks. The principal's decisions (offers) will generally depend on her assessment of the agent's motivation: his self–confidence in his ability to perform the task, and his beliefs about its difficulty or the personal payoffs (intrinsic rewards) attached to it. Consequently, the agent has an incentive to strategically represent his self–view to the principal: he may for instance signal high self–confidence when interviewing for a job or a new position, or low self–confidence when aspiring to an easy–going assignment.

[35] At least in terms of point–estimates and uncertainty. Ideally, one would also like to independently measure their true underlying preferences (susceptibility to temptation). This is likely to be quite difficult, however, precisely because the individual may be following a general rule which involves mimicking someone with higher willpower.

2) The "*looking–glass self.*"[36] Conversely, the individual's social environment often attempts to manipulate his self–view, or (very similarly) his perceptions of current and future activities. Thus, parents and educators try to boost children's confidence in their ability ("you can do it"), point at the large delayed payoff attached to a good education, or understate the unpleasantness of tasks ("math is fun"). Such strategies require that the principal hold information that is not available to the agent. For example a teacher, thesis advisor, coach or therapist may, due to her background and prior experience, be better able to judge the talent and prospects of success of a pupil, student, player, or patient. A parent is likely to have better information than their child about the monetary and nonmonetary payoffs to education, and also be better at evaluating their achievements.[37] In all these situations, it is not just the principal's words (encouragement, praise, blame) but also the kinds of incentives (rewards, punishments) which she offers to the agent that provide the latter with a "looking glass" from which he can learn about himself.

Both self–presentation and the looking–glass self are well amenable to standard economic analysis, based on signalling theory. In the rest of this section we shall focus on the latter situation, which is more original from an economic point of view, and also more directly related to this chapter's central concern with self–knowledge and motivation. As before, the agent's self–view is being manipulated in order to affect his performance, but the manipulator is now external (the principal) rather than internal (a previous incarnation). Consequently, the modeling of the looking–glass self and its implications requires only one of the three grains of sand introduced in Section II, namely imperfect self–knowledge.[38]

The key point in the analysis is then that the agent views the principal's behavior as reflecting her information about him, and conversely the principal realizes that her words and deeds will lead the agent to revise his beliefs and alter his choices. For this looking–glass effect to operate, however, it is not sufficient that the principal have private information relevant to the agent's decision–making. As explained in the next section, two other conditions are required:

(i) *Sorting.* The principal's incentives to offer different types of rewards or feedback must vary (preferably, monotonically) with what she knows about the agent or the task to be performed.

(ii) *Attribution.* The agent must understand –at least approximately– how the principal's motives impact her behavior.[39]

[36] This term is borrowed from Cooley (1902).

[37] Alternatively, she may be trying to convey information about her own preferences, ("I really want you to do well in school"), which in turn could be selfish ("make us proud") or altruistic ("I really care about your future").

[38] Of course, having the other two operate as well would only enrich the set of issues that can be addressed. For instance, one important role for altruistic principals (parents, teachers) could be to help the agent overcome his time–consistency problem; but of course, they also have their own agendas (e.g., morals, pride, reputation).

[39] Psychologists have accumulated a substantial amount of evidence on the variety and relative sophistication of inferences made by individuals in social environments. See, e.g., Festinger (1954), Heider (1958), Kelley (1976), or Gilbert and Silvera (1996).

V.1 Intrinsic versus extrinsic motivation

In Bénabou and Tirole (1999b) we study the provision of incentives (rewards, empowerment, feedback) in educational and workplace environments. A principal (parent, teacher, manager) possesses information relevant to the agent's (child, student, subordinate) decision–making, and the agent tries to see through the principal's ulterior motivation when responding to the incentive scheme he is offered.

Most of economics is built upon the premise that people respond to incentives, and there is a good amount of evidence that they usually do. In other words, rewards serve as "positive reinforcers" for the desired behavior. In psychology, their effect is more controversial. Psychologists, while not denying that incentives often work (except for some who argue that rewards are alienating per se), point to a number of experiments and real–life situations to argue that they are only weak reinforcers in the short term, and negative reinforcers once they are withdrawn.[40] That is, rewards have hidden costs, which economists typically neglect.

An old–style behaviorist approach, mapping in reduced form the stimulus (reward) into a response (reduced effort) is unlikely to be productive, as it would say little as to how the response is altered with changes in the environment. For instance, how does a reward offered to a child for passing an exam differ from stock options held by a start–up entrepreneur? A cognitive / economic modeling of the question takes a more structural approach, reflecting the simple intuition that contingent rewards may "send the wrong signal". This is, for instance, why most parents realize that offering toys, candy or even hugs and kisses to their child for each time they read a book, complete their homework or clean up their room is a usually bad idea. The analysis thus brings to light the two key factors that help us understand when rewards are positive or negative reinforcers: the extent to which the principal holds *private information*, and the direction of her *"sorting condition"*. The first insight suggests for instance that one should expect performance rewards to have (on average) greater hidden costs in an educational environment than in the workplace. Indeed, a young person is typically quite uncertain about his self and the nature of the tasks that he faces, while the parent or educator can make judgements based on previous experience with other children or students. In the workplace, by contrast, both the nature of tasks and the reward scheme are usually less individual–specific and more publicly known, being part of a "job description" or relatively standard contract.

Let us now illustrate the crucial role of the second insight, namely the sorting condition. Suppose first that the principal feels less of a need to offer a reward to the agent when she trusts in his ability. This may arise for two reasons. First, rewards may be more costly if they are handed out often, i.e., if the agent is likely to succeed. Second, the principal may have private information but still be unsure about the agent's precise ability and/or motivation. A principal who has every

[40] See, e.g. Deci (1975), Wilson et al. (1981), Kruglanski et al. (1971), Lepper et al. (1973), or Kohn (1993). Recent evidence by economists includes Gächter and Fehr (1998) and Gneezy and Rustichini (2001).

reason to believe that the agent should feel good about himself or like his assignment need not provide much of a reward; conversely, if she is worried that the agent is or will become discouraged about his chances, she will need to offer him stronger incentives. However, if the agent, in turn, tries to read through the principal's choice of reward what she knows about him or his assignment, then an offer of reward represents bad news, and therefore has an hidden cost on the motivation side. It can in fact be shown in such settings that: a) rewards are weak positive reinforcers in the current task; b) once withdrawn, a reward indeed becomes a negative reinforcer.

Result (a) means that the direct incentive effect of rewards must dominate the negative inferential effect –otherwise rewards would be self–defeating and would not be offered by rational principals –except by mistake, or in experiments. Result (b) allows us to makes economic sense of the conflict between "intrinsic" and "extrinsic" motivation emphasized by a huge literature in psychology and sociology (e.g., Deci 1975, Deci and Ryan 1985). At the same time, however, it delineates *empirically testable* limits to its validity. In particular, we cannot stress too much the importance of the sorting condition. Indeed, one can think of cases where rewards are positive reinforcers even in the long term (i.e. once withdrawn). For example, suppose that effort in the current task allows the agent to "learn by doing", and that this learning by doing is particularly effective for more talented agents. The sorting condition then operates in the opposite direction: the principal is more eager to reward, the more talented the agent. As a result, rewards are now interpreted as good news. Similarly, the principal may be signalling how much she cares about the agent's success in the task; if the agent somewhat internalizes the principal's utility, this generates a positive response of effort to reward. Yet another illustration is offered by the issue of help where, depending on the situation, a principal may have more of an incentive to lend a "helping hand" when the agent is weak, or when he is strong. For example, some Ph.D. supervisors spend substantial time with the weakest students (who will have trouble finishing their thesis in time) and with the strongest ones (in whose reflected glory they may bask, or whose work they find intellectually most pleasing). It is therefore not too surprising that help sometimes creates a pattern of dependency, while in other situations it is received as encouraging news.

V.2 Undermining the other's ego

While boosting or avoiding damage to others' self–confidence is a pervasive aspect of social interactions, people also often criticize or downplay the achievements of their spouse, child, colleague, subordinate or teammate, and disclose information that is detrimental to their egos. In our (1999b) paper we therefore also address this kind of behavior, and show that it is not inconsistent with the need (stressed in the social psychology and human resource management literatures) to coach and boost the self–esteem of one's personal and professional partners. There are at least three rational motives for such ego–bashing: a) the two agents may be in competition; b) the other may be tempted to "rest on his laurels"; c) more interestingly, the two may be engaged in a battle for dominance. Indeed, when decision–making relies on an agreement between two

parties, as is often the case in dyads, each may be tempted to demonstrate that he or she has a better judgement than the other (in the matter at hand, or in general), in order to impose his or her views and choices. The very fact that individuals attempt both to boost heir partners' self–esteem at certain times, and to deflate it at others, demonstrates once again the need for an explicit, structural approach to the description of social interactions.

VI Concluding Comments

Economics has always been about making a small number of strong assumptions to allow for broad predictive power and policy analysis. At the same time, economists have been eager to extend their paradigm beyond its traditional realm in order to incorporate the lessons from other social sciences such as psychology, sociology, law or political science. In particular, the limits of the "Homo Economicus" paradigm are well–known, and isolated pioneers –economists or psychologists working across boundaries– have long tried to broaden its scope. In doing so they paved the way for the more systematic effort at cross–fertilization undertaken in recent years by both scientific communities, and reflected in particular by the present volume.

The value of this enterprise for economists is clear, and goes well beyond the "private benefits" of entering new and intellectually challenging territory. We can surely expect to improve our understanding of numerous fields such as consumer behavior, product design, organizations, education, health or finance. The real challenge will be to find a proper balance between two methodological extremes: an ad hoc, hypothesis–intensive approach on the one hand, and a conservative (or compulsive?) attachment to rationality and functionalism on the other. The present chapter has tried to demonstrate how introducing a small number of key imperfections into the standard paradigm of Economic Man can go a long way towards reconciling it with a large body of disparate, and sometimes apparently conflicting, evidence from psychology. Needless to say, our work only scratches the surface; much remains to be done and–as evidenced by the contributions in his volume– there is room for many alternative approaches.

The value of the joint enterprise for psychologists is of course not for us to judge. We can only hope that they will find enough in it for them to sustain a lasting dialogue, and at this point we can only speculate on what some of these potential benefits could be (aside from the intellectual satisfaction of seeing economists finally pay proper attention to their findings). First, there might be a revival of interest in developing a motivated and relatively unified approach to human behavior, bringing greater coherence to what sometimes seems like a collection of unrelated –sometimes even contradictory– effects and biases. Second, new models will lead to new predictions and therefore, down the line, to new experiments. We sketched several potential directions in this chapter, but in practice there is likely to be a lag while formal theory first struggles to satisfactorily incorporate the large corpus of preexisting evidence.

References

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

[2] Ainslie (2001) *Breakdown of Will,* in press. Cambridge University Press.

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

[4] Akerlof, G. and Kramton, R. (2000) "Economics and Identity," *Quarterly Journal of Economics,* 115(3), 715–753.

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

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

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

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

[9] Baumeister (2001) "The Psychology of Irrationality: Why People Make Foolish, Self-Defeating Choices," this volume.

[10] Baumeister, R., Heatherton, T. and Tice, D. (1994) *Losing Control: How and Why People Fail at Self–Regulation.* Academic Press: San Diego, CA.

[11] Bénabou, R. and J. Tirole (1999a) "Self–Confidence: Intrapersonal Strategies." IDEI mimeo, June.

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

[13] Bénabou, R. and J. Tirole (2000) "Willpower and personal Rules," Princeton University mimeo, June.

[14] Berglas, S. and Baumeister, R. (1993) "Your Own Worst Enemy: Understanding the Paradox of Self–Defeating Behavior." BasicBooks: New York.

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

[16] Bodner, R. and Prelec, D. (2000) "A Neo–Calvinist Model of Conscience," this volume.

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

[18] Brocas, I. and Carrillo, J. (2000) "Information and Self–Control," this volume.

[19] Caillaud, B., Cohen, D. and Jullien, B. (1999) "Towards a Theory of Self–Restraint," CERAS mimeo, December.

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

[21] Carrillo, J., (1998) "Self Control, Moderate Consumption, and Craving," CEPR D.P. 2017, November.

[22] Carrillo, J., and T. Mariotti (2000) "Strategic Ignorance as a Self–Disciplining Device," *Review of Economic Studies,* 67(3), 529–544.

[23] Cooley, C. (1902) *Human Nature and the Social Order*, New York: Scribner's.

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

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

[26] Deci, E., and R. Ryan (1985) *Intrinsic Motivation and Self–Determination in Human Behavior*, New York: Plenum Press.

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

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

[29] Gabaix, X. and Laibson, D.(2000) "Bounded Rationality and Directed Cognition," Harvard University mimeo.

[30] Gächter, S. and Fehr, E. (1998) "How Effective are Trust- and Reciprocity-Based Incentives? in: A. Ben-Ner and L. Putterman (eds): *Economics Values and Organizations,* Cambridge University.

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

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

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

[34] Gneezy, U. and Rustichini (2001) "Pay Enough or Don't Pay at All", Quarterly Journal of Economics, forthcoming.

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

[36] Gul, F. and Pesendorfer, W. (1999) *"An Economic Theory of Self–Control,"* Econometrica, forthcoming.

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

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

[39] Herrnstein. R., Loewenstein, G., Prelec, D. and Vaughan, W. (1993) "Utility Maximization and Melioration: Internalities in Individual Choice," *Journal of Behavioral Decision Making,* 6, 149–185.

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

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

[42] Kahneman, D., Wakker, P. and Sarin, R. (1997) "Back to Bentham? Explorations of Experienced Utility," *Quarterly Journal of Economics*, 112(2), 375–407.

[43] Kahneman, D. (2001) "Experienced Utility and Objective Happiness: a Moment–Based Approach," this volume.

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

[45] Kohn, A. (1993) *Punished by Rewards*, New York: Plenum Press.

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

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

[48] Kruglanski, A., Friedman, I. and G. Zeevi (1971) "The Effect of Extrinsic Incentives on Some Qualitative Aspects of Task Performance," *Journal of Personality*, 39: 608–617.

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

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

[51] Lepper, M., Greene, D., and R. Nisbett (1973) "Undermining Children's Interest with Extrinsic Rewards: A Test of the 'Overjustification Hypothesis'," *Journal of Personality and Social Psychology*, 28: 129–137.

[52] Loewenstein, G. (1996) "Out of Control: Visceral Influences in Behavior," *Organizational Behavior and Human Decision Processes,* 65(3) March, 272–292.

[53] Loewenstein, G. and Schkade, D. (1999) "Wouldn't It Be Nice? Predicting Future Feelings," in *Well–being: Foundations of Hedonic Psychology,'* D. Kahneman, E. Diener and N. Schwartz, eds. New York, NY: Russel Sage Foundation

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

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

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

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

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

[59] Piccione, M. and Rubinstein, A. (2000) "On the Interpretation of Decision Problems with Imperfect Recall," *Games and Economic Behavior, 20, 3–24.*

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

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

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

[63] Seligman, M. (1975) *"Helplessness: On Depression, Development, and Death." an Francisco, CA: Freeman and Co.*

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

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

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

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

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

[69] Thaler, R.H. (2000) "From Homo Economicus to Homo Sapiens," *Journal of Economic Perspectives, 14(1),* Winter, 114–142.

[70] Thaler, R.H. and Sheffrin, H. M. (1981). *"An Economic Theory of Self Control,"* Journal of Political Economy, 89,2, pp. 392–406

[71] Weinberg, B. (1999) "A Model of Overconfidence," Ohio State University mimeo, August.

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

[73] Wilson, T., Hull, J. and J. Johnson (1981) "Awareness and Self–Perception: Verbal Reports on Internal States," *Journal of Personality and Social Psychology*, 40: 53–71.

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