Essays in Behavioral Economics

by

Botond Kőszegi

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Economics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2000

Author .....
Department of Economics
April 14, 2000

Certified by .....
Peter A. Diamond
Institute Professor
Thesis Supervisor

Certified by .....
Jonathan Gruber
Professor of Economics
Thesis Supervisor

Accepted by .....
Peter Temin
Elisha Gray II Professor of Economics; Chairman, Departmental
Committee on Graduate Students

Essays in Behavioral Economics

by

Botond Kőszegi

Submitted to the Department of Economics
on April 14, 2000, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics

Abstract

Chapter 1: This chapter examines the logical consequences of the rather unsurprising notion that humans care about and manage their self-image, a notion long taken for granted by psychologists. I model this by assuming that decisionmakers derive utility from positive views about the self, holding constant standard utilitarian outcomes usually assumed relevant in economics. Other than this, agents are time-consistent expected utility maximizers, are constrained in their updating by Bayes' rule, and can manipulate their beliefs only by controlling the flow of information that they receive. The motive to maintain a favorable self-image leads to a systematic rejection of free information about the self in certain states of the world and eventually to overconfident beliefs. Economically relevant decisions are affected by this overconfidence as well as the incentive to gather information about and make decisions so as to optimally manage beliefs. Agents might avoid informative actions when satisfied with their current beliefs ('self-image protection'), and seek out activities in which they can prove themselves when they are not ('self-image enhancement'), even if these choices are otherwise poor. These motives lead to a whole host of effects on behavior that other models have trouble explaining in a unified framework. The model can also make testable predictions on how these effects play themselves out across different categories of tasks and within a category of tasks over time. Applications to stock market participation, the choice between salaried and self-employment, career choice, manager behavior, and employee motivation are discussed.

Chapter 2: This chapter starts from the same premise as the previous one, the assumption that agents care about their self-image, but examines its consequences in a different information structure. Agents can improve financial decisions by making subjective judgments about their payoffs, while they derive ego utility from their perceptions regarding this ability. If the agent has a self-image protection motive, she will as a result be averse to making a subjective judgment and reviewing it later, since this combination is informative about ability. The consequence could be a sluggishness in responding to new information, procrastination in making up one's mind, or the reliance on inferior objective information. Possible remedies and applications are discussed, with particular attention to anxiety about health.

Chapter 3 (with Peter Diamond): There is overwhelming psychological evidence that some people run into self-control problems regularly, yet the effect of these findings on major life-cycle decisions hasn't been studied in detail. This paper extends Laibson's quasi-hyperbolic discounting savings model, in which each intertemporal self realizes that her time discount structure will lead to preference changes, and thus plays a game with her future selves. By making retirement endogenous, savings affect both consumption and work in the future. From earlier selves' points of view, the deciding self tends to retire too early, so it is possible that the self before saves less to induce her to work. However, still earlier selves think the pre-retirement self may do this too much, leading to possible higher saving on their part and eventual early retirement. Thus, the consumption path exhibits observational non-equivalence with exponential discounting. Observational non-equivalence also obtains on a number of comparative statics questions. For example, a self could have a negative marginal propensity to consume out of changes in future income. The outcome with naive agents, who fail to realize their self-control problem, is also briefly discussed. In that case, the deciding self's potential decision to retire despite earlier selves' plans results in a downward updating of available lifetime resources, and an empirically observed downward jump in the consumption path.

Thesis Supervisor: Peter A. Diamond
Title: Institute Professor

Thesis Supervisor: Jonathan Gruber
Title: Professor of Economics

Acknowledgments

MIT has been a wonderful learning environment. It was a joy to talk to Peter Diamond from the start. His guidance has shaped my interests and taught me the basics of turning ideas into models and using models to sharpen ideas. It has also been a pleasure to work with Jon Gruber: as a result, I had to realize empirical work can be a lot of fun. And I'll never forget the unrelenting maximalism of Sendhil Mullainathan, who pushed me towards something more interesting every time we had a conversation.

But it was my friends and family who made the last few years a really good time. I hope to have convinced my parents that going into economics wasn't such a bad idea. My sister Bogi has been a friend like only a sister can be. The movies and beers with my bro Bulcsú had been my most common escapes out of here. Among my colleagues at MIT, I would especially like to thank Max Amarante, Saku Aura, Kobi Braude, Jon Levin, Markus Möbius, and Kathy Yuan. I hope our paths will cross many times in the future.

And, my favorite rabbit–I love you.

Contents

Chapter 1

Ego Utility and Overconfidence

1.1 Introduction

Contact with information plays a central role in organizing or motivating most kinds of economic behavior. When we invest in stocks, try smoking, apply for jobs, decide to attend school, choose whether to go for physicals, or make almost any purchase, we take advantage of a wide array of information in the form of published data, our experiences and those of acquaintances, subjective judgments, and introspective evidence on our tastes, abilities, and moods.

In most work on information within economics, decisionmakers receive and precisely interpret externally given signals they care about for the sole purpose of choosing an action they will take. Psychologists have long recognized two problems with this modeling standard: humans commit errors in processing information, and they care more strongly about the conclusions they reach based on the information. The focus of the current paper is the second criticism, which is now supported by an overwhelming array of evidence: in experiments, certain kinds of information elicit important affective, even physiological, reactions that subjects anticipate and care about, while in surveys, it's clear that people want to view their traits, abilities, and choices in a positive way. Furthermore, information gathering is influenced by these desires in addition to the usefulness and cost of information–it seems that people are willing to go a long way to protect their favorable beliefs about the self.

Motivated by this psychological evidence, by observation, and by introspection, I study the 'hot' aspects of information processing in a simple rational model in which beliefs enter the utility function directly: agents derive pleasure from positive views about the self, holding constant standard utilitarian outcomes usually assumed relevant in economics. For example, an older woman might enjoy being reminded that she was the prettiest girl on her prom night, although this is now unlikely to benefit her in any way. The model used here assumes that individuals' behavior is motivated both by this 'ego utility' and the traditional, 'instrumental' side of utility. Ego utility is tightly linked to an actual choice problem–agents find a trait desirable if it qualifies them for a socially desirable action. Specifically, they like to think they

are good enough to engage in the more ambitious of two activities, one in which only higher types can perform well. I examine the consequences of these preferences for agents who are limited in generating positive beliefs about themselves by Bayes' rule, without making the claim that all these consequences are exclusively rational in their origin. Rather, the key reason for using the Bayesian approach is methodological: by changing just one assumption, in this case adding a new dimension of utility, we can isolate the behavioral effects of that single assumption. Here, agents motivated by ego utility, and whose only means of manipulating their beliefs is controlling the flow of information that they receive, exhibit a whole host of behaviors traditional models would have difficulty explaining in a unified setting.

The first logical consequence of my setup is a *level* effect on beliefs derived from information acquisition about the self *before* the agent enters the choice situation in question: through learning, people become overconfident in their ability to perform the more ambitious task. I define overconfidence as describing the situation where, if asked, too many agents would (rationally and honestly) report that they think they would make more money with the ambitious option. This results from the fact that agents *want* to *think* that this is the case, so as they voluntarily collect information about their ability, they are more likely to stop when they think it is. This is quite a general result that is very hard to eliminate despite the Bayesian context.

The rest of the results of the basic model are driven by the agent's desire to affect *marginal* changes in her beliefs, that is, to manipulate the choice of tasks to manage the signals she receives about herself. There are two reasons why the agent may choose to do this. Firstly, when she is satisfied with her present beliefs, she might distort[1] her instrumental choices to avoid receiving information about herself. I call this the *self-image protection* motive. On the other hand, if she is dissatisfied with her current perception of herself, she might go out of her way to try to improve her

[1]The word 'distort' might be confusing here, since this is a rational model of an agent acting in isolation, who thus behaves optimally given her preferences. The predicted behaviors only look like distortions to someone who ignores ego utility. Throughout the paper, this term refers to differences on the *instrumental* side of behavior relative to what would be expected from an agent with no ego utility.

beliefs. This is the *self-image enhancement* motive.

In addition to predicting overconfidence, self-image protection, and self-image enhancement as the basic consequences of ego utility, the model also organizes how these effects play themselves out across different categories of tasks and within a category of tasks over time. Specifically, the behavioral distortions caused by self-image protection and self-image enhancement depend on two *observable* features of the choice situation: the relative informativeness of the ambitious and unambitious tasks and on the opportunities to learn about ability before entering.

In the basic model, agents have ample opportunity to learn in advance. Also, if the ambitious and unambitious tasks are equally informative about ability, then the self-image protection and enhancement motives do not result in a distortion in the choice of task. The agent does what she thinks is best financially based on her current beliefs, and so the overconfidence in beliefs translates directly into overconfidence in observed actions. However, for many applications the more natural assumption is that the ambitious task is more informative about ability. In this case, self-image protection may lead the agent to choose the unambitious task and self-image enhancement might make her choose the ambitious one. Specifically, when she thinks she is good but she is not quite sure, the decisionmaker will avoid the ambitious option for fear of embarrassing herself. Thus, the self-image protection motive can couple overconfidence in beliefs with underconfidence in observed actions! On the other hand, an agent who tries the ambitious task and is disappointed by herself might attempt to prove herself by trying it again–this is the self-image enhancement motive. Self-image protection can affect agents at each stage of their participation in the task, while self-image enhancement is particular to those who have been in for a longer time; since agents can learn about their ability before entering, they will start the activity with favorable beliefs about themselves, and only during the course of making decisions (possibly) get disappointed, endogenously acquiring a desire to reassert themselves.

The prediction that the agent enters the choice situation with protective tendencies and only later could switch to self-image enhancement is weakened when there is

little opportunity to learn beforehand or when learning can only be done using very accurate information relative to that available through the task. In both cases, pre-choice learning is limited, so the agent doesn't have the opportunity to manipulate her beliefs in a positive direction before entering, therefore having to use the activity to try to do so. All these predictions make the model falsifiable, even though beliefs are not directly observed by an econometrician [2],[3].

The abstract model in this paper is intended to be a relatively general model of distortions in behavior due to self-image or pride, with many potential economic applications. Consider two examples: an investor's choice of participating in the stock market or investing in a safe option like a checking account, and an actor's decision between movies and television. One needs to be a good enough investor to do well in the stock market, and similarly, it is harder to survive in moviemaking than in tv. I would argue that picking stocks is more informative about one's ability as an investor than having a bank account. On the other hand, one can learn just as much about one's acting skill in television as in moviemaking. If decisionmakers attach ego utility to thinking that they are good enough to be on the lucrative side of investing or acting, the model predicts overconfidence: when asked, too many people will claim they would be successful *if they chose the ambitious goal.* In the acting example, people put their money where their mouth is: they actually try movies too often, leading to overparticipation and a lot of disappointments. However, self-image protection prevents some decisionmakers with positive beliefs from entering the stock market, so under some conditions too few actually enter. As a result of self-image

[2]In this sense, this model resembles many others we use which rely on unobservable variables. See, for example, efficiency wage theories (Shapiro and Stiglitz 1984), Jovanovic's (1982) theory of small firm growth, or various models of intergenerational altruism. Analogously to beliefs about the self in my model of self-image utility, a worker's shirking decision, a firm's underlying permanent efficiency parameter, or parents' altruistic feelings are not directly observable. Just as these models make predictions about inter-industry wage differentials (Krueger and Summers 1988), the growth and failure of enterprises, and consumption and the allocation of income (Altonji, Hayashi, and Kotlikoff 1992), respectively, I can make empirical predictions about the relationship of variables that we do observe.

[3]One interpretation of this model is one in which beliefs are understood as a consumption good– they enter the utility function just as other goods do. Even though the consumption of beliefs is not directly observable, it is indirectly observable because information is an input into beliefs. This gives the model its falsifiability.

enhancement, investors who enter and lose might stay in too long, particularly by holding on to their losing stocks.

Other applications include small businesses, project choice by managers, career or educational choice, job search, and the relationship between extrinsic and intrinsic motivation of employees. Some of these examples are discussed in section 1.5.

1.2 Psychological Evidence

This section briefly reviews some evidence indicating the sensibility of the paper's non-standard assumption–that agents have preferences over their beliefs about themselves. The first part focuses on an indirect prediction of ego utility: that people hold incredibly positive views of their traits and prospects. Although there are multiple ways to explain this fact, I will suggest that the most parsimonious way is to assume that humans *prefer to* believe these nice things, even at the cost of being inaccurate. The second part of this section discusses more direct psychological evidence on the assumption of ego utility itself. Unfortunately, psychologists have not tested this proposition directly, so none of the evidence is conclusive. However, the sum-total of all the evidence, especially including the biases in the first part, is highly suggestive.

For those readers who are not interested in details of the psychology literature, this section can be skipped without loss of continuity.

It is common knowledge that people hold rather favorable views of their abilities. In an instance of this that might be economically relevant, Meyer (1975) surveyed General Electric Company employees about their self-ratings on a percentile scale compared to others in similar jobs and with similar salaries. Over 90% of those surveyed rated themselves as above average. The average rating was at the 77th percentile. Only two out of about a 100 surveyed rated themselves below average–both putting themselves at the 45th percentile! Similarly, Svenson (1981) reports that 90% of us consider themselves better than the median driver. Biases are also found in reports of intelligence (Wells and Sweeney 1986), leadership ability, taking good notes in class, and memory (Campbell 1986). People are also more likely to attribute

positive than negative outcomes to the self (e.g. Taylor and Brown (1988)) and to believe that their first judgments are right (e.g. Lord, Ross, and Lepper (1979)) [4].

It is unlikely that people report these views purely for self-presentational purposes, as some suggest. The biases are not responsive to manipulations about anonymity or whether the subjects believe their lies can be detected (Greenwald and Breckler 1985) [5]. In addition, some studies indicate that subjects will act on biased beliefs [6].

Unrealistic views are not limited to ability. In general, people seem to be overoptimistic about their prospects. They estimate the likelihood that they will experience a variety of pleasant events, including getting a good salary, having a gifted child, living past 80, doing well in an investment, and traveling, to be higher than for their peers (Weinstein 1980). On the other hand, they believe themselves to be less likely than others to have an automobile accident, to be a crime victim, being depressed, contracting various illnesses, and even having gum problems (Perloff and Fetzer 1986) [7].

That many of these biases tend to take time to develop (Burger (1986), Burger and Huntzinger (1985)) testifies to the notion that there is something nontrivially cognitive going on. Also, the biases are responsive to incentives, and are stronger under conditions of ambiguity (Brown 1986), indicating that they are limited by rationality at least to some degree. In as much as these beliefs are actually rational, I would argue that the most sensible reason they overwhelmingly go in the direction of what people would like to see happen or what is socially desirable is exactly that people prefer to believe socially desirable things about themselves. Other reasons consistent with Bayesian updating that have been proposed in the economics literature potentially go a certain distance in explaining the biases, but they can't explain findings of the above generality.

---

[4] These are called the self-serving and confirmatory biases, respectively.

[5] With the so-called bogus pipeline manipulation, most subjects can be made to believe that the lie detector they are connected to works reliably. This manipulation decreases biases somewhat, but the biases don't disappear (Riess, Rosenfeld, Melburg, and Tedeschi 1981).

[6] An example is Staw's (1974) study that will be discussed below.

[7] Two exceptions are smoking-related illnesses and AIDS. People grossly overestimate the negative effects of smoking (Viscusi 1994), and the probability that a single intercourse will lead to the transmission of HIV.

The most notable set of these alternative explanations is centered around self-control problems [8]. A decisionmaker who is aware of her problem (a 'sophisticated hyperbolic discounter') recognizes that she will not behave properly in the future, and also that the information she collects affects not only her current but her future choices as well. If certain posterior beliefs are more likely to induce 'correct' behavior in the future, the current self wants to manipulate her information collection strategy so that she is more likely to end up with these posteriors. For example, if one has a self-control problem regarding unprotected sex, it might be beneficial to believe that such behavior leads very easily to contracting HIV, so that one is kept in control by this 'fear.' Therefore, the agent stops looking for evidence on HIV when she thinks it is easily transmitted (Carillo and Mariotti 1997). Similar effects are explored in Benabou and Tirole (1999b) and Carillo (1997). But it is unlikely that hyperbolic discounting can explain all the biases and overoptimism we have mentioned. If believing that HIV can strike anytime is good to control our desire to engage in casual irresponsible sex, then why isn't this the case for our susceptibility to cancer and eating habits? Also, if ability and effort are substitutes in some task (as opposed to complements as in Benabou and Tirole (1999b)), one would rather have negative views about the self to induce effort in the future [9].

Of course, the biases about prospects, abilities, and traits are probably partly irrational, not solely a consequence of Bayesian updating. People might systematically deceive themselves by dismissing negative information and taking positive information very seriously. In this case, it is impossible to rule out purely cognitive reasons; for example, if people expect to succeed and are mentally prepared for favorable outcomes, it may well be easier to process information about success, so that good news get a greater weight in updating. But I would still argue that the cognitive mistakes

[8] See also Zábojnik (1999), whose model is driven by the type-dependent cost of information collection.

[9] Turning around the message of the papers by Carillo (1997), Carillo and Mariotti (1997), and Benabou and Tirole (1999b), one can understand the above examples as situations in which preferences over beliefs can help overcome self-control problems. In other words, positive beliefs don't arise from self-control problems, but self-control problems are alleviated by positive beliefs rooted in an intrinsic preference for them. Self-control problems are then a possible evolutionary explanation for these beliefs.

in question are also motivated–why would people generally expect to succeed? Or, if the mistakes are understood in a multiple self context, with one self fooling the other (an implicit interpretation of some non-Bayesian models), why would the former self want to deceive the latter? Also, cognitive explanations of some experiments where (alleged) motivational factors are manipulated are quite far-fetched (Kunda 1987). Whether or not biases are exacerbated by cognitive errors, it seems very natural to conjecture that people just like believing good things about themselves. Since this is the premise of the current paper, we turn to the more direct psychological evidence that it is the case below.

The most direct evidence that people have preferences over what they think about themselves comes from the extensive literature on cognitive dissonance. Cognitive dissonance theory, originally proposed by Festinger (1957), but greatly refined since, postulates that people can't hold two contradictory cognitions for long, and so when they do, one of the cognitions has to disappear. I am interested in a specific experimental design for studying cognitive dissonance: inducing people to do something they consider wrong. This is conjectured to bring about an unpleasant state called dissonance arousal, which usually leads subjects to change their opinion about the act in question in order to reduce that arousal [10]. For our purposes, the important prediction is the first part (dissonance arousal), and the second (attitude change) is only a diagnostic tool to identify the presence of the first. Based on the evidence, it seems reasonable to conclude that dissonance arousal comes from a negative judgment about the self, which the subject then tries to relieve by changing her opinion.

Dissonance arousal has measurable physiological consequences related to stress. These include the frequency of spontaneous electrodermal activity (Croyle and Cooper 1983) and the constriction of blood vessels in the outer portions of the body (Gerard (1967), Petty and Cacioppo (1981)). In addition, the dissonance manipulation facilitates performance on simple tasks but interferes with performance on complex tasks (Pallak and Pitman (1972), Gaes, Melburg, and Tedeschi (1986)). This means

[10]For example, subjects who wrote pro-marijuana essays for children feel bad and thus convince themselves that marijuana use is not so bad after all.

it energizes the *dominant responses*, which is known to be an expression of arousal (Spence, Farber, and McFann (1956), Cooper and Fazio (1984)). In short, having done something aversive has physical effects, and thus a source of utility in the most basic sense.

By ingeniously manipulating the amount of stress or its attribution to various sources, experiments by Zanna and Cooper (1976) and Cooper, Zanna, and Taves (1975) attempt to demonstrate that the state of dissonance arousal is a necessary mediator for attitude change to occur [11]. This indicates that attitude change is a response, and a partially conscious one, to the state of dissonance arousal, a conclusion that allows researchers to use the indirect measure of attitude change as an indicator of arousal–a very simple method for the laboratory.

Cooper and Fazio (1984) summarize the necessary features of a choice situation that can be expected to produce dissonance arousal. Their list is consistent with the notion that dissonance arousal occurs when the (induced) behavior allows the subject to make a negative judgment about herself [12]. First, the choice must lead to an aversive event or at least the possibility of an aversive event. If subjects make a counterattitudinal speech on the legalization of marijuana to a non-committed audience, they experience greater attitude change than when the audience is firmly committed in either direction (Nel, Helmreich, and Aronson 1969). Similarly, if people believe their audience remains unconvinced, they don't show attitude change (Cooper and Worchel 1970). Second, the choice must be to at least some degree irrevocable. In

[11] Zanna and Cooper (1976) gave subjects placebo before the usual dissonance manipulation, but convinced them that the drug was real with either no side effects or possible side effects of 'tenseness' or 'relaxation.' People in the 'tense' condition changed their attitudes least, and those in the 'relaxation' condition most–apparently because in the former case subjects were able to misattribute their tension from dissonant behavior to the drug. In a flip side of this experiment, Cooper, Zanna, and Taves (1975) gave people either a tranquilizer, placebo, or amphetamine prior to the dissonance manipulation, but made them believe they were all ingesting placebo. As expected, those who took amphetamine changed their attitudes most, and those who took the tranquilizer least.

[12] One could directly test whether a negative judgment about the self is an ingredient of cognitive dissonance arousal. A simple design is running the high choice/low choice conditions, but with an extra wrinkle: for some participants, making the outcome entirely irreflexive of the subject by, for example, making the connection random. I venture to guess that dissonance arousal would be much lower with random connections, indicating that judgments about the self, and not only regret, is what drives arousal.

Davis and Jones' (Davis and Jones 1960) experiment, some subjects but not others were lead to believe they would be able to correct for their dissonant action. Those who thought they would have this opportunity exhibited less change in attitude. Third, the agent must perceive an element of personal responsibility in the aversive outcome. Freedom of choice is certainly an ingredient of responsibility: in all experiments on this topic, subjects who think they can choose not to perform the undesirable action, but still do so, change their attitudes more than those who perceive less freedom. In addition, the unintended outcome must be foreseeable–in order for subjects to experience dissonance arousal, they must be able to tell themselves they 'should have seen it coming' (Goethals, Cooper, and Naficy 1979).

Once again, it is not very likely that reported attitude change is only a self-presentational device. Staw's study analyzes a natural experiment in the ROTC, which many young men joined partially to avoid being drafted. Consequently receiving a high draft lottery number puts this avoidance strategy in a less favorable light. Among those who have committed to join the ROTC, those who received a high draft number had more favorable views of the ROTC, presumably because they had to justify their choice. These people also did better in the program (Staw 1974). It is hard to imagine how such behavior would come about unless the beliefs are internalized.

Cognitive dissonance theory also has indirect predictions on selective information processing. However, experiments in this area are weakly formulated, and don't take the rational/Bayesian framework seriously enough [13]. It is no surprise that many predictions have met with rather shaky empirical support (Frey (1986), Frey and Wicklund (1978)). There is a lot more evidence, however, that people often make an effort to avoid free information about the self. If utility didn't have an ego dimension, this would in general not be the case [14]. For example, in Brock and Balloun (1967),

[13]For example, it is often argued that dissonance arousal can be reduced by selectively looking at decision-consonant information and avoiding contradictory information (Frey 1986). In a rational framework, the decisionmaker might reduce dissonance by either looking carefully at positive information or by trying to discredit negative information–there is no a priori asymmetry.

[14]although there are a few very specific cases where people should avoid free information. For example, most of us wouldn't want to know the end of a thriller before we see it. Or, many scientists don't want to know the work of Nazi doctors who used prisoners as experimental subjects. In the examples to be presented, the novelty of information doesn't carry value in itself, and there is no

religious people and smokers prefer to obscure messages that run counter to their commitments. More to the point, many patients choose to avoid genetic information on their susceptibility to breast cancer (Lermna et al (1998)). Also, in Larrick and Boles' (1995) experiment, subjects in a negotiation setting behave so as to avoid feedback on the unchosen alternative [15].

The above array of evidence demonstrates quite forcefully that judgments about the self on the basis of past choices create changes in people's internal states (to use a term as general as possible) that they might want to control. More generally, there is evidence that people are affected by their beliefs about various abilities.

Unfortunately, though theorists often take it for granted (see e.g. Harter (1985) for a review), evidence that agents' attributions about their ability affect their mood/affective/arous states is rare, possibly because those attributions are more difficult to manipulate. A study by McFarland and Ross (1982) managed to do so through the manipulation of feedback on a fake social accuracy test. According to the subsequent mood questionnaire, subjects felt better when they succeeded than when they failed only if they attributed the result to ability. Other evidence from our daily lives, though not establishing a causal link, is nevertheless suggestive. We all feel good about ourselves, for example, when we receive praise at graduation or play well in a soccer game. What I'm talking about, of course, is pride. Psychologists describe pride as "an emotional response to an evaluation of one's competence" or as a "self-reward" (the internal counterpart of praise) (Lea and Webley 1997), both indicating a highly non-neutral state.

Finally, most forms of self-handicapping are understood by psychologists as information-avoiding behavior driven by the desire to protect good judgments about the self. Setting goals that are unreachable or structuring a situation to be excessively difficult, procrastination, overcommitment, and perhaps even substance abuse are ways to avoid attributions of low ability after failures (Fiske and Taylor 1991). Thus, these

obvious 'principle' which should make subjects avoid it.

[15]Unfortunately, most experiments in the psychology literature let subjects choose *between* pieces of information rather than *whether or not* to see the information.

behaviors constitute an expensive way of suppressing the informativeness of a failure outcome in a task. And research indicates that at least part of the motive to engage in self-handicapping is to maintain a high self-esteem (Arkin and Baumgardner 1985).

1.3 A Model of Ambition

1.3.1 Setup

This section develops a basic model of choice grounded in a simple premise: that people have pride about what activity they see themselves capable of succeeding in. To illustrate the key results, I use a parsimonious reduced-form model driven by a specific structure on the ego utility function that is compatible with this starting point. Some finer details of the decision problem are addressed in sections 2.2 and 2.3.

The decision of investors to stay in bonds or go into stocks is a useful example to illustrate the intuition. I will use the language of this example to describe the results, but the model is a more general one with multiple applications, some of which I will discuss in section 1.5. It takes a better investor to do well in the stock market, but the returns are potentially also higher, so it makes sense to enter if one is good enough [16]. In addition to caring about financial returns, people are assumed to have a degree of pride about this ability–they derive pleasure from thinking that they belong to the distinguished class of people who are likely to make money in the stock market. Another application with the same formal structure is the decision between having a salaried job and starting one's own business, with the latter presumably being more challenging.

Formally, the model has two conceptually different kinds of activities, seperated into different decision periods: an initial learning period (period 0) in which the agent can gather information about herself, and subsequent choice periods which

[16]The ability I'm talking about is not necessarily restricted to picking stocks, however: it could partly be a skill of avoiding being cheated by the broker, or one of matching investments to one's needs.

directly affect financial outcomes. For ease of exposition, we will gradually increase the number of financially relevant decision periods from zero to two.

In each of the choice periods, the agent can choose between two options, with the options being the same over time. Option 1 (bonds) is riskless and unambitious, while option 2 (stocks) is risky and ambitious. The riskless option has a certain payoff of 0 in each period, while the risky option can give either a payoff of -1 or a payoff of 1. For simplicity, we assume that the agent is risk-neutral with respect to financial payoffs–the predictions regarding the behavioral consequences of self-image would be unchanged by the addition of risk aversion.

The risky option is called ambitious because its expected payoff is higher for agents with higher ability: whether it leads to success (earns a payoff of 1) depends on the agent's underlying ability q in addition to a random luck variable. Specifically, a signal $s^t = q + \epsilon^t$ is generated independently in each period t according to $\epsilon^t \sim N(0, \sigma_s^2)$, and the ambitious undertaking succeeds in that period if $s^t > 0$. Thus, one needs to be either good or lucky to do well in the stock market, but success is more likely when one is good. We will consider various assumptions on the choices that lead to the observation of s^t: this signal might be observed after only one choice or after both. The distribution of q in the population is $N(0, \sigma^2)$, and this is also the prior the agent starts off with.

In the initial learning period, period 0, the agent can collect information about the ability parameter that affects her success in the stock market. Specifically, learning takes the following (somewhat ad hoc) form. The agent can observe an arbitrary, possibly infinite, number of signals $s^{0j} = q + \epsilon^{0j}$, where $\epsilon^{0j} \sim N(0, \sigma_s^2)$ and these errors are independent. These signals are only available before she enters the actual financial decision problem; she will not be able to see them again.

This assumption is meant to capture the idea that people enter different situations having a notion about how well they will do in them, but how much and what they know depends at least partially on themselves. The specific assumptions I have made–that the agent can learn her type perfectly, but can also completely stop learning, and if she does, she can't resume later–are rather extreme in this regard, but they

eliminate technical problems without hurting the basic points.

Timing is crucial in this problem, since utility depends directly on the amount of information received. Thus, not only the timing of decisions, but also the timing of the realization of ego utility is important. When there are no financial decisions to be made, the agent simply samples signals s^{0j} in period 0 until she decides to stop (if she does), and then her ego utility is realized in period 0'. With one financial choice period, the timing is the following:

0: sequentially observe signals s^{0j} and, after each one, decide whether or not to stop

1: choose whether to be in bonds or stocks

1': possibly observe s^1

1": ego utility realized

As we have mentioned, we will consider various assumptions about the conditions under which s^1 is observed. When there is one more period, the last three stages are repeated in the same order.

The agent is a Bayesian information processor and likes to think she is good enough to try the ambitious option, investing in the stock market. Precisely, her ego utility is one if, according to her current beliefs, it makes financial sense to go for option 2; it is zero otherwise. Notice that–since the agent is risk-neutral in terms of financial payoffs–this is the case when the subjective probability of getting the high outcome 1 is at least a half. Since the subjective distribution of s^t is centered around the mean of current beliefs about ability, that probability is greater than one-half iff the mean is greater than zero. u, then, depends only on the mean of the agent's

beliefs, and in period $t = 0, 1, 2$

$$u(\bar{q}(S^t)) = \begin{cases} 1 & \text{if } \bar{q}(S^t) > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{1.1}$$

where S^t is the set of signals observed up to the end of period t and $\bar{q}(S)$ is the agent's mean belief conditional on her prior and the set of signals S [17].

The agent is an expected utility maximizer with von Neumann-Morgenstern utility function in periods $t = 1, 2$ of [18]

$$wu(\bar{q}(S^t)) + a_i^t, \tag{1.2}$$

where a_i^t is the outcome of the option chosen in period t. w is a weighting parameter.

1.3.2 Why a Bayesian Expected Utility Model?

The modeling approach in this paper is rather unique: it uses a Bayesian model to capture phenomena many regard and model as non-Bayesian (Akerlof and Dickens (1982), Rabin and Schrag (1999), Taylor and Brown (1988)), and it keeps expected utility but abandons the substitution axiom at a crucial point. Both of these choices deserve a few words.

As mentioned in the introduction, I believe that many of the phenomena I will discuss, and especially the overconfidence result, are at least in part non-Bayesian in their origin. However, I also believe that the self-image motivation is the driving force behind the phenomena I will discuss (it even induces us to make, in its favor, many cognitive mistakes I will not discuss), so following standard theoretical methodology, I'm changing just one assumption relative to existing models to see how far that assumption by itself will take us. The Bayesian model also has the advantage

[17] Obviously, we've made enough assumptions for the agent's beliefs to stay normal after each update.

[18] In Baumeister's (1998) language, ego utility would correspond to the 'reflexive self,' the part of our psyche that examines ourselves. Instrumental utility motivates the 'executive self,' which is responsible for making decisions. My model is then about a decision problem where both of these selves are evoked at the same time.

over existing non-Bayesian or quasi-Bayesian ones [19] in that it carries easily across contexts; it is not specific to the information structure given. In addition, I will argue in section 1.4.5 that the Bayesian model captures much of the flavor of a non-Bayesian model driven by self-image utility, with few disadvantages.

Besides a methodologically attractive starting point, there is another way to understand the Bayesian model. An implicit assumption of the Bayesian approach is that the agent can't lie to herself about her ability. For example, she can't choose to tell herself that she is a good investor, feel euphoric for a while, and then choose to realize that she is not so good and keep her money in a checking account. It's clear that humans are capable of suspending disbelief–we do so every time we get emotional about a movie. However, our ability to reward ourselves by deliberate self-delusion is limited (Heider 1958). The model explores the behavior of an agent acting under this limitation by making the extreme assumption of Bayesian updating.

There is also an apparent tension in my formulation of the agent's objective function, in that she is assumed to be an expected utility maximizer with von Neumann-Morgenstern utility function as in 1.2, yet the substitution axiom of expected utlity is abandoned through the non-linearity of u in beliefs. However, this assumption is well justified on normative grounds.

Recall the substitution axiom: if the lottery L is preferred over the lottery L', then for any lottery L'' and $\alpha \in (0,1)$, the mixed lottery $\alpha L + (1-\alpha)L''$ is preferred over $\alpha L' + (1-\alpha)L''$. The intuitive appeal of this axiom comes from a non-complementarity argument: one can think of the mixed lotteries as compound lotteries, in which the agent gets (with probability $1-\alpha$) L'' or (with probability α) L and L', respectively. Since in a realized state of the world it should be irrelevant what would have happened in another state (the agent will never experience that state), she should choose the compound lottery lottery featuring L rather than L'.

This argument hinges on a limited understanding of what payoffs decisionmakers care about. It assumes that payoffs are only realized when the state of the world

[19] For example, see Rabin and Schrag (1999), Akerlof and Dickens (1982), or Gervais and Odean (1999).

has been revealed. In particular, if decisionmakers derive utility *directly* from their beliefs, not only through what they imply for other payoffs, the above intuition loses its appeal: since different states of the world are included in beliefs together, complementarities can easily arise. And in talking about self-image, it is natural to assume that people derive utility from their *beliefs* about their ability, not from their *actual* ability, the latter of which they don't know. Similarly, the above argument fails for any hedonic experience that depends on expectations about the future [20], since those can only depend on beliefs about the future as opposed to what actually happens. On the other hand, once utility from beliefs is added to the space of admissible payoffs, the appeal of the substitution axiom comes back–when all utility-relevant outcomes are taken into account, non-complementarity in compound lotteries over those outcomes should once again hold, even if it doesn't hold for compound lotteries over physical payoffs.

1.4 Results

Fact 1 *Starting from beliefs* $N(\bar{q}_0, \sigma_0^2)$ *and after observing signals* $v^1, v^2, \ldots, v^n$ *distributed according to* $N(q, \sigma_s^2)$*, the posterior mean of the agent's beliefs is*

$$\frac{\sigma_s^2}{\sigma_s^2 + n\sigma_0^2}\bar{q}_0 + \frac{\sigma_0^2}{\sigma_s^2 + n\sigma_0^2}\sum_{i=1}^{n} v^i, \tag{1.3}$$

and the corresponding posterior variance is

$$\frac{\sigma_s^2\sigma_0^2}{\sigma_s^2 + n\sigma_0^2}. \tag{1.4}$$

This implies that the variance of the agent's posterior mean is

$$\frac{n\sigma_0^2}{\sigma_s^2 + n\sigma_0^2}\sigma_0^2. \tag{1.5}$$

[20] such as anxiety, which will be discussed in section 2.7 of chapter 2.

From this fact, we will use the following two things extensively in the current section. First, the variance of the posterior only depends on the number of signals observed, so we can use the notation $\sigma_p^2(J)$ for the variance of the agent's posterior after having observed J signals. And second, the variance of the posterior mean is increasing in the number of signals observed.

We start with establishing existence of optimal strategies. For this, we restrict the agent's stopping rules in the learning period to be measurable with respect to the measure generating the sequence of signals.

Theorem 1 *An optimal strategy exists.*

Proof. See appendix. □

The description of the optimal strategy is complex for other than the most trivial cases. However, the sections below establish the properties we need for the more interesting results.

1.4.1 Overconfidence in Beliefs

We start with the simplest case, when there is only an initial learning period, period 0–there are no financial outcomes, and so the only relevant part of the agent's utility function is ego utility. In this case, since she can only lose by sampling further, the agent will stop updating whenever the mean of her current beliefs is greater than zero. Conversely, she will continue updating when her current mean beliefs are less than or equal to zero.

Fact 2 *Suppose there is no financially relevant decision period. After observing the signals* $s^{01}, \ldots, s^{0J}$*, the agent chooses to observe the next signal if and only if* $\bar{q}(\{s^{01}, \ldots, s^{0J}\}) \leq 0$.

The final beliefs thus have the property that all agents with below-zero beliefs know their types for sure, while this is not the case for those with good beliefs. As a consequence, some of those holding good beliefs are in fact bad types (with actual

$q \leq 0$), and not vice versa, so more people will be holding good beliefs than there are good types.

If we conduct a survey in this population asking people 'Do you think you would make more money in the stock market than in the bond market?', more people will answer 'Yes.' than is actually the case [21]. Therefore, even though agents are Bayesian and so unbiased in their beliefs about the underlying ability parameter q, they *are* biased in the following very reasonable confidence measure:

Confidence measure: Proportion of agents reporting mean beliefs greater than zero at the end of period 0; equivalently, $p_{conf} = Prob(\bar{q}(S^0) > 0)$.

Actual ability measure: Proportion of agents with actual ability greater than zero; equivalently, $p_{act} = Prob(q > 0)$.

As we have seen, 'pride' or ego utility resulted in $p_{conf} > p_{act}$. This point is a very general one that nevertheless has largely been unappreciated in the literature. We receive information in all kinds of forms, and based on our priors and the information obtained, we make various conclusions about specific abilities and prospects that we care about. If the function that maps beliefs to conclusions is not linear, we can't apply the law of iterated expectations to the space we care about, so we can't in general expect the conclusions to match what we expect from the population distribution of abilities. And then, if we want to believe in desirable conclusions in the relevant space, we will be biased [22]. And once again, we can arrive at this bias

[21] This question really only makes sense when there are financially relevant decision periods, but let's ignore that–the issues to be discussed are not limited to this first model, but they are easier to express here.

[22] In our example on ambition, the space where the signals are received and where the ability parameter is drawn from is $\Re$, the set of real numbers. The relevant space for conclusions is a two-point set, an answer to the question 'Am I good enough for the stock market?' A non-constant

with only one tool to control our beliefs: the ability to stop information flow at will.

As presented here, the overconfidence result seems vulnerable to the exact form of the confidence measure used. In particular, if we asked the question 'Is your expected excess return in the stock market at least $\frac{1}{2}$?', we might get underconfidence: if the signals are inaccurate (σ_s^2 is large), too few people will answer this question affirmatively. However, I would argue that in practice it is very difficult to ask such a question, because that would require the surveyor to know a lot about the ability space on which respondents base their judgments. For most surveys, the scale on which the question is asked is one that people attach ego utility to, so we should get a positive bias as in this model. We should expect unbiased answers only if we asked people for their probability distribution function of the parameter in question, since this question forces linearity on the structure. Importantly, even this prediction depends strongly on the assumption of correct and identical priors. If agents have dispersed but on average correct priors, the overconfidence result comes back even for probability distribution functions: agents with too low priors will tend to seek out information more than agents with too high priors, increasing average mean beliefs relative to the priors.

I still do not claim, based on the above, that overconfidence is the result of completely rational information processing, but the Bayesian model goes surprisingly far in this respect [23]. In addition, it generates some interesting specific predictions on the form of beliefs that do not easily come out of a non-Bayesian model. For example, in this model the confident agents (those holding positive beliefs) are overconfident

map between normal distributions over $\Re$ and $\{0,1\}$ is *necessarily* non-linear.

[23] I'm not the first one to make the point that Bayesian updating can lead to systematic biases in certain measures–this point has been made by Carillo and Mariotti (1997) and Carillo (1997) in the context of hyperbolic discounting (see section 1.2), by Zábojnik (1999) in the context of costly learning with type-dependent costs, and by Rabin (1995) in the context of moral constraints and self-serving biases. As far as I know, I am the first one to make the connection between self-image and overconfidence, to say that some biases are likely to be very general even in the Bayesian context, and to study the distinction and interplay between confidence in beliefs and confidence in observed actions. In some sense, the difference is a matter of focus: while the former papers focus on generating preferences for information acquisition through instrumental considerations, I take these preferences as given in the ego utility function, and examine its consequences. Also, partly because of my focus on self-image, this model also has a much broader applicability than the others mentioned.

on average (some of them are in fact not of high ability), while the less confident are accurate. There is some evidence in the psychology literature that beliefs about the self have this property (e.g. Coyne and Gotlib (1983)). Also, the current setup predicts that agents would rate themselves more favorably relative to how others *are* than relative to how others *rate themselves.* In essence, this is because they would answer the questions 'How many people do you think would make more money in the stock market?' and 'How many people do you think would claim they would make more money in the stock market?' very differently. There is very little evidence on whether this prediction holds, but subjects' post facto justifications of self-other discrepancies in ability judgments in Campbell's (1986) study point in that direction.

1.4.2 Overconfidence in Actions

By adding financially relevant decision periods, we now allow for both ego and instrumental utility to influence the agent's behavior. We assume first that s^t, the signal upon which success depends if the ambitious option is chosen, is observed after either choice. Thus, even after choosing the unambitious option, the agent finds out how she would have done had she chosen the ambitious one.

This section shows that the overconfidence result holds even in the presence of financial incentives to find out more about one's ability. Theorem 2 proves that, not surprisingly, people who have a pessimistic self-image will be eager to find out new things about themselves in period 0–in general, this helps them both in becoming happier about the self and making better financial decisions. However, a similar statement is not true for those who are happy about their abilities. Theorem 3 shows that some of these people will stop acquiring information about themselves. The two results together imply overconfidence in the previous measure (based on the number of agents who believe they would make money in the stock market), although it will be clear that the presence of financial incentives reduces the degree of overconfidence.

First, note that *conditional on* her information, the agent always chooses the financially sensible option; that is, she chooses to invest in stocks if and only if that yields a higher return in expectation, or, when the mean of her current beliefs is

greater than zero. This is simply because the options only differ in their financial payoff–they don't differ in their informativeness, so they affect ego utility in the same way.

Theorem 2 *Suppose there are one or two financially relevant decision periods and that s^1 and s^2 are observed no matter which option is chosen. After observing the signals $s^{01}, \ldots, s^{0J}$, the agent chooses to observe the next signal in period 0 if $\bar{q}(\{s^{01}, \ldots, s^{0J}\}) \leq 0$.*

Proof. We show that for any beliefs of the above kind, it is better to see at least one more signal than stopping. By fact 1, the agent is more likely to believe that she is of a high type in each of the two periods if she observes the next signal, so her expected ego utility will be higher. On the other hand, she can do at least as well financially, since she can follow the exact same strategy as she would have otherwise. □

Theorem 3 *Suppose there are one or two financially relevant decision periods and that s^1 and s^2 are observed no matter which option is chosen. For any $w > 0$ and positive J, there is a positive integer $J' \geq J$ and a nonzero measure of signals $s^{01}, \ldots, s^{0J'}$ with $\bar{q}(\{s^{01}, \ldots, s^{0J'}\}) > 0$ after which the agent chooses not to observe the next signal in period 0.*

Proof. Suppose by contradiction that for almost all beliefs the agent chooses to continue drawing signals if at least J signals have been observed; that is, she almost always chooses to find out her type. We will prove that in this case there is a non-zero measure of posterior beliefs where it is better to stop. More specifically, for the given posterior variance $\sigma_p^2(J)$, for a sufficiently high $\bar{q}(\{s^{01}, \ldots, s^{0J}\})$, the agent will choose not to sample further. For simplicity we use the shorthands $\bar{q} = \bar{q}(\{s^{01}, \ldots, s^{0J}\})$ and $\sigma_p^2 = \sigma_p^2(J)$ for this proof.

By sampling on, the agent will with probability $\Phi\left(\frac{-\bar{q}}{\sigma_p}\right)$ find out that she is of low type. Take $\bar{q}$ to be high enough so that if she stops drawing signals, the probability that her posterior mean is moved below zero in the one or two periods of choice is

negligible compared to $\Phi\left(\frac{-\bar{q}}{\sigma_p}\right)$ [24]. This implies that compared to this probability, the difference between the financial payoffs of the agent's actual strategy and the strategy of choosing option 2 in both periods is negligible, and her expected ego utility differs from w in a relatively negligible way as well.

If she samples, the agent can change her choice to option 1 in those states of the world where this is more sensible financially. Her payoff is then 0 instead of $1 - 2\Phi\left(\frac{-q}{\sigma_s}\right)$, giving a savings of $f(q) = 2\Phi\left(\frac{-q}{\sigma_s}\right) - 1$. Total savings relative to always choosing option 2, and ignoring the above negligible possibility that the agent chooses option 1 in period 2 without finding out her type, is then

$$\int_{-\infty}^{0} \frac{1}{\sigma_p}\phi\left(\frac{q-\bar{q}}{\sigma_p}\right) f(q)dq. \tag{1.6}$$

Now notice that the function f is continuous, bounded, decreasing, and satisfies $f(0) = 0$. Since, as $\bar{q}$ becomes large, the distribution $\frac{1}{\sigma_p}\phi\left(\frac{q-\bar{q}}{\sigma_p}\right)$ becomes arbitrarily right-weighted, the above integral becomes arbitrarily small compared to $\Phi\left(\frac{-\bar{q}}{\sigma_p}\right)$. Thus, eventually the financial gains from updating will be outweighed by the utility losses. □

It's clear that a more general version of this theorem, in which there are more than two periods of decision, would hold as well.

Recall that since the two tasks are equivalent in informativeness, the agent in this model chooses what she thinks is best financially. And the overconfidence just derived makes her think the stock market is better too often, so she will choose to participate too often. Therefore, the overconfidence in beliefs is translated into overconfidence in observed actions.

In this population, many will try the ambitious option for a while and find out that they are not good enough for it, learning this 'the hard way' instead of from the free signals at the beginning. The failure rate, which we can define as the percentage of agents experimenting with the ambitious option who later give up, and which we

[24] Such a $\bar{q}$ exists because (by fact 1) the variance of the agent's posterior mean is smaller than the variance of her current beliefs.

would normatively expect to be zero in this problem if we ignored ego utility, will consequently be positive. On the other hand, there won't be any 'success stories:' no one will start with the unambitious option and later move to the ambitious one [25]. In other contexts, these aggregate observations might lead one to conclude that agents are *irrationally* confident, but here they arise purely as a result of people's desire to hold on to their positive beliefs, if only for a while.

1.4.3 Self-Image Protection...

We now change the assumption that s^t, the signal on which success depends if the ambitious option is chosen, can be observed with either choice. Instead, we suppose that in order to see whether one would succeed in an activity, one has to *try it.* For many applications, including stock market participation, this is a more realistic assumption. We start with the case when there is one financially relevant decision period.

Theorems 2 and 3 still go through with similar proofs. The new result is self-image protection.

Theorem 4 *Assume there is one financially relevant decision period and that s^1 is observed if and only if the ambitious option is chosen in period 1. Suppose further that $S^0 = \{s^{01}, \ldots, s^{0J}\}$. Then there is a unique $q^*(J) > 0$ such that the agent chooses the ambitious option if and only if $\bar{q}(S^0) > q^*(J)$.*

Proof. Since the number of signals is finite, from theorem 2 we know that $\bar{q}(S^0) > 0$. For $\bar{q}(S^0) \approx 0$, the agent is approximately indifferent financially between the ambitious and unambitious options. At the same time, in terms of ego utility she strictly prefers the unambitious option, since by engaging in the ambitious activity she might be forced to update her beliefs down to the negative side. This implies that for $\bar{q}(S^0)$ close enough to zero, she chooses the unambitious option.

[25] This population behaves the same way as could be expected from a population of slow learners with inflated priors.

As $\bar{q}(S^0)$ increases, choosing the ambitious option becomes strictly more attractive financially (the probability of getting a payoff of 1 increases) and in terms of ego utility (the probability that beliefs will drop to below zero decreases). And for a sufficiently high $\bar{q}(S^0)$, when beliefs almost certainly stay positive after observing s^1, the agent wants to choose the ambitious option. □

Essentially the same proof will be used in the next section to establish the self-image enhancement motive, that some agents with uncertain negative mean beliefs will choose the ambitious activity. The reason we don't get that here is not that the proof wouldn't work for a one-period model, but that in a one-period model no one starts the activity with uncertain negative beliefs: anyone who has negative beliefs has already observed an infinite number of signals in period 0.

Self-image protection arises when the agent has good, but uncertain, views of herself. Since she is afraid of finding bad news, she'd rather not take on something challenging (in this case investing in the stock market), even though the challenging choice seems reasonable by financial standards.

In the area of personal investment, people often make the apparent mistake that they invest in inferior vehicles, putting their money in checking accounts, for example, when clearly better alternatives are available at minimal switching costs (financial or otherwise). This could be due to an irrational time-inconsistency-induced procrastination as in O'Donoghue and Rabin (1999b): the agent keeps putting off the decision because she doesn't want to bear the small cost of doing it, but always thinks that soon she will. However, in this setting such behavior arises naturally in a rational manner–if more profitable investment involves difficult judgments, people might want to engage in them because they are afraid they would feel stupid if they failed.

The model with a self-image protection motive shares the property of the model without it (sections 1.4.1 and 1.4.2) that more people think they are good enough for the ambitious option than would be predicted from the population distribution of abilities. The curious situation here, however, is that these beliefs are not necessarily translated into action, creating a sort of rationally grounded hypocracy. In fact, people's actions and beliefs diverge for exactly the same reason that their beliefs

are inflated: they want these beliefs to be good. Depending on the parameters, the proportion of investors entering the stock market could be greater than or less than one-half, the proportion whose financial expected returns are actually higher if they invest in stocks. Therefore, we can get a situation in which most people realistically expect financial profits from the ambitious option, yet few of them actually choose to take advantage of it. For a non-financial example, note that many researchers, though they think of themselves as very talented, shy away from really difficult research projects.

Relative participation in the ambitious task depending on which tasks lead to an observation of s^t provides a clean, testable, out-of-sample prediction of the model. Let us focus on small businesses. If one could find appropriate measures for the key elements of the model, a neat test would be looking at participation in ambitious businesses depending on when the uncertainty about business success is resolved. If it's resolved early, then the ambitious choice is more informative about business skill. On the other hand, if success is resolved late, the ambitious and unambitious tasks are more equally informative; the interpretation of this case in terms of the current model would be that s^t is *not* observed with either choice, which clearly has implications similar to the case when s^t is observed with both. The prediction of the model would be that there is more overparticipation and failure when success is resolved late. Of course, businesses with different resolution times have different risk characteristics: if it turns out early that one wouldn't succeed, one can just get out. Therefore, an empirical project has to control carefully for risk aversion. But if this is done properly, one also gets an idea of the relative importance of risk aversion and self-image protection in people's tendency to shy away from risky and ability-sensitive adventures [26].

[26] I believe the term 'risk aversion' is applied too broadly in most economic discussions. It can denote reference-free financial risk aversion, financial risk aversion mediated by a reference point (loss aversion), or information aversion disguised as financial risk aversion. Consider the stock market example: the uncertainty in stock returns is not only financial risk, but also information about investing skill, since it provides information about the quality of investment judgments (see also section 2.2 of chapter 2). Investors might be averse to this uncertainty for both financial risk-aversion and self-image protection reasons, and self-image protection can potentially explain part of the equity premium puzzle not accounted for by standard calibrations of risk aversion.

1.4.4 ... and Enhancement

Finally, we add one more financial decision period, while keeping the assumption that s^t is observed if and only if the ambitious option is chosen in period t. Theorems 2 and 3 still hold, but the proofs are somewhat more complicated.

Theorem 2 Version 2 *Assume that there are two financially relevant decision periods and that s^t is observed if and only if the ambitious option is chosen in period t. After observing the signals $s^{01}, \ldots, s^{0J}$, the agent chooses to observe the next signal in period 0 if $\bar{q}(\{s^{01}, \ldots, s^{0J}\}) \leq 0$.*

Proof. The principle of the proof is similar to that of theorem 2, but there is one complication arising from the fact that now the agent can manipulate her financial options in order to change beliefs. We distinguish two cases:

Case 1. If the agent stops information acquisition, she chooses the ambitious option in the first period.

In this case, it's clear that it's better to instead draw one more signal and then choose the unambitious option: that provides the same signal (and thus the same effect on ego utility) and is better financially.

Case 2. If the agent stops information acquisition, she chooses the unambitious option in the first period.

Then, it's better to draw one more signal and make the same financial choices as otherwise. □

Theorem 3 Version 2 *Assume that there are two financially relevant decision periods and that s^t is observed if and only if the ambitious option is chosen in period t. For any $w > 0$ and positive J, there is a positive integer $J' \geq J$ and a nonzero measure of signals $s^{01}, \ldots, s^{0J'}$ with $\bar{q}(\{s^{01}, \ldots, s^{0J}\}) > 0$ after which the agent chooses not to observe the next signal in period 0.*

Proof. The proof is similar to that of theorem 3, but there is one wrinkle. Since the unambitious choice is now not informative, the agent can manipulate her choices to try to increase her ego utility at the expense of financial payoffs. However, it is

clear that the per-period gain from doing this is at most w times the probability that two signals will get her down to the zero ego utility level. But this probability, as we saw in the proof of theorem 3, is, for a large enough $\bar{q}$, dwarfed arbitrarily by the probability of getting there through a perfect learning of type. □

The main purpose of adding an extra period to the decision problem is the following theorem. It proves that in the second period, a self-image enhancement motive can distort financial choices in addition to self-image protection.

Theorem 5 *Assume that there are two financially relevant decision periods and that s^t is observed if and only if the ambitious option is chosen in period t. Suppose further that $S^1 = \{s^{01}, \ldots, s^{0J}\}$ or $S^1 = \{s^{01}, \ldots, s^{0J-1}, s^1\}$. Then there is a unique $q^*(J) > 0$ such that the agent chooses the ambitious option in period 2 if and only if $0 \geq \bar{q}(S^1) > -q^*(J)$ or $\bar{q}(S^1) > q^*(J)$.*

Proof. For $\bar{q}(S^1) > 0$, the self-image protection motive can be established in the same way as in theorem 4. The cutoff mean belief value q' at the agent is indifferent between the options satisfies

$$1 - 2\Phi\left(\frac{-q'}{\sqrt{\sigma_p^2(J) + \sigma_s^2}}\right) - w\Phi\left(\frac{-q'}{\sqrt{\frac{\sigma_p^2(J)}{\sigma_p^2(J)+\sigma_s^2}\sigma_p^2(J)}}\right) = 0. \tag{1.7}$$

Now consider $\bar{q}(S^1) < 0$. Symmetrically to the other case, for $\bar{q}(S^1) \approx 0$, the agent is approximately indifferent financially, so she will choose the ambitious option, since this might put her back on the good beliefs side on ego utility. As $\bar{q}(S^1)$ decreases, choosing the ambitious option becomes worse both financially (the probability of getting a payoff of 1 decreases) and in terms of ego utility (the probability that the agent will end up getting her beliefs high again decreases.) The cutoff value q'' at which the agent is indifferent between the two options satisfies

$$1 - 2\Phi\left(\frac{-q''}{\sqrt{\sigma_p^2(J) + \sigma_s^2}}\right) + w\left(1 - \Phi\left(\frac{-q''}{\sqrt{\frac{\sigma_p^2(J)}{\sigma_p^2(J)+\sigma_s^2}\sigma_p^2(J)}}\right)\right) = 0 \tag{1.8}$$

It is easy to prove that $q' = -q''$. $\square$

The new behavioral distortion (relative to expected wealth maximization) is called self-image enhancement. It is the possibility that the agent tries to prove herself by engaging in the challenging activity despite probable losses. She does this hoping that things will turn out well in the end, and she can feel positive about herself again. Indeed, there is at least some evidence that when investors lose in the stock market, they tend to stay in too long, particularly by holding on to their losing investments (Odean 1998). We will return to this issue in section 2.3 of chapter 2.

To complete our discussion of the basic effects in the model, we prove that the self-image protection motive in the first period carries over in a qualitatively similar way from before: the agent might choose the unambitious option with positive mean beliefs, but only if the beliefs are close enough to zero. This highlights a key distinction between self-image protection through ignorance of free signals and through distorting financial choices. If we see an agent refusing to consider free information, we can't in general conclude that she is of a relatively low type–in fact, in some cases the agents who ignore free signals are the highest types. However, if an agent distorts her financial behavior relative to her beliefs, as with self-image protection, we know those beliefs are not very secure.

Theorem 6 *Assume that there are two financially relevant decision periods and that s^t is observed if and only if the ambitious option is chosen in period t. Suppose further that $S^0 = \{s^{01}, \ldots, s^{0J}\}$. Then there is a unique $q^{**}(J) \geq 0$ such that the agent chooses the ambitious option in period 1 if and only if $\bar{q}(S^0) > q^{**}(J)$.*

Proof. Once again we prove that choosing the ambitious option becomes better in differential terms as $\bar{q}(S^0)$ increases. There are two cases.

Case 1. If the agent is unambitious in the first period, she will be ambitious in the second.

In this case, the agent's expected first-period utility if ambitious is equal to her expected second-period utility if unambitious. Therefore, the difference between choosing the ambitious versus unambitious option in the first period depends on the ex-

pected second-period utility with the ambitious choice. And clearly, this increases with $\bar{q}(S^0)$.

Case 2. If the agent is unambitious in the first period, she will be unambitious in the second.

This case is even easier: the agent's utility is constant for the unambitious option, while it clearly increases with $\bar{q}(S^0)$ for the ambitious one. □

Note: although we have not proven that $q^{**}(J)$ is greater than zero, we know that for $\bar{q}(S^0) \approx 0$, the agent either i. did not stop drawing signals with these beliefs in the initial learning period, and so can't have them upon choice; or ii. she will choose the unambitious option. The reason is that if it's worthwhile to choose the ambitious option, then it must have been worthwhile to draw one more signal–doing so and then choosing the unambitious option is just as good, and in some states of the world it is better.

A motive for self-image protection can thus arise in either of the two periods, while self-image enhancement can only happen in the second. This is a consequence of our assumption that the agent can view an arbitrary number of signals about ability before she enters the financial decision problem. If she wants to enhance her beliefs about herself at the beginning, she can do so by just observing another signal; she does not need to distort her financial behavior. On the other hand, if she is disappointed in her attempt at the ambitious option, in the second period the only way to regain her pride is by trying it again. Thus, people whose egos are threatened are the ones who take extra risks to prove themselves [27].

Although the above model sharpens this distinction to its extreme, the basic intuition is likely to hold in many settings. We have all kinds of opportunities to learn about ourselves in the course of everyday life, and when we choose not to learn any longer, it is likely to be when we don't want to, that is, when we are on the concave parts of our ego utility functions. This means that we are likely to enter an average new situation averse to new information about ourselves, leading to

[27] America's need to prove itself through extra risk taking is documented in Time magazine's September 6, 1999 cover story on the new trends of thrill seeking.

information-jamming behavior by, for example, going for safe options. It is during the course of doing the new activity that we might move to the convex part of u, and we want to reestablish a satisfactory self-image by looking for extra signals through the manipulation of task choices.

Of course, there are also dimensions of ability in which cheap learning opportunities are available at any time, eliminating the self-image enhancement motive. In the basic model, if the agent can sample an arbitrary number of free signals in each of the periods (not only in period 0), she never wants to distort her financial choices to improve her beliefs–she would rather sample free signals. However, she might still distort her choice of tasks to protect her self-image. The difference is fundamental: while it is possible to substitute toward different signals when the agent wants information, this is impossible in the case of protection. Besides the general result that people enter an average situation with protective tendencies, this constitutes an extra reason why we should expect to see more self-image protection in the real world.

While unreasonable in the stock market context, there might also be applications in which the ambitious option is *less* informative about the relevant ego parameter than the unambitious one. In this case, self-image protection and enhancement reinforce the financial decisions beliefs would already justify. When the agent thinks instrumental considerations warrant the ambitious option, self-image protection gives her an extra reason to choose it. Conversely, when she has bad beliefs, she wants to be unambitious for both instrumental and ego reasons. Therefore, beliefs and actions coincide once again. The difference relative to section 1.4.2 is that there is no region of approximate indifference between the two options: the agent discretely prefers either one or the other. Also, overconfident beliefs are very persistent in this setting, because once established, there is no chance for the agent to be proven wrong.

Examples in which the ambitious option is less informative are probably less common than those where it is equally or more informative. Brandenburger and Nalebuff (1997) note that some people believe they are indispensable for their firm, and thus never take a vacation. Of course, this way they never find out if they are really so important. Another example is the spying dilemma of parents, for instance about

their kids' drug use: parents might prefer to believe that 'things are not so bad' to warrant spying, whereas spying is the activity through which this can be verified [28]. The perhaps uncomfortable prediction of my model is that parents spy too little, hoping to avoid the issue. At the same time, the preference shift from not spying to spying is discrete, even if induced by a small piece of information.

An interesting complication I have avoided in the current model is related to the *fine-tuning of beliefs*. To see this, consider a slight modification of the model: in the initial updating period, suppose that there is only one signal available, and after that the agent can only jump to finding out her type exactly or not receive any more signals. Somewhat surprisingly, with this modification it is no longer in general true that agents with negative mean beliefs want to find out their type. If, say, w is so high that the agent practically only cares about ego utility, and her mean beliefs are close to zero, she definitely won't take the perfect signal: by instead first engaging in the ambitious activity and then conditioning her next 'move' on the outcome, she can push her beliefs above zero with a higher probability.

What drives the agent's behavior is her fear of finality, or, in economics terminology, her option value for ego utility. Even though finding out her type can only make her feel better in terms of self-image, this is only true for today: if she finds out she is not good, she will not only still feel bad, but know for sure that she will *always* feel bad. That's why she prefers an information structure with which she can manipulate her beliefs more easily. In general, holding other things like total informativeness and cost equal, agents who care about ego utility will prefer information that comes in smaller pieces. Versions of this principle are stated formally in the following fact.

Fact 3 Option value

1. *Suppose the only relevant part of the agent's utility function is the ego utility u, and consider any u. If, from the point of view of period 0, the agent is indifferent whether to find out her type, she prefers not to find it out.*

[28] Of course, the ego parameter in this case has no convenient interpretation in terms of ability. But it is likely that parents attach self-image utility to what their children are like.

2. *Suppose that in period 1, the agent can observe a signal about her ability that becomes available again in period 2. If the period 1 instrumental and ego utility values of the signal are zero, the agent prefers not to observe the signal in period 1.*

Therefore, when the information associated with the tasks has less accurate signals than the ones available in pre-choice learning, agents might voluntarily choose not to learn very much. They might also be involuntarily prevented from learning if there is little opportunity to learn about themselves outside the tasks. In either case, they aren't able to manipulate their beliefs before the actual choice problem, so it is not in general true that they will enter with a self-image protection motive. This provides another testable prediction of the model based on the opportunities to learn before instrumentally relevant behavior is observed. In addition, when pre-choice learning is voluntarily or involuntarily limited, if overconfidence develops, it does so during performance of the tasks. This means that average confidence increases at the beginning of the activity, while exactly the opposite should be the case in the first-pass model of this section.

The signal structure of the model in the current section is designed to avoid a potential complication from the standard option value concern: if the agent doesn't know her type, she might choose the more informative option even if it doesn't make financial sense in light of the current period, because by doing so she finds out valuable information about herself. This I have eliminated by making sure that if the agent has uncertain negative mean beliefs when choosing, she is in the second period, so that the option value is not an issue. This makes the definition of ego utility much simpler [29].

[29]This option value issue is distinct from the option value for ego utility discussed above. In fact, the distinction is quite interesting: whereas the standard option value consideration would induce one to take more informative actions, option value for ego utility would make one take less informative ones, all else equal.

One more caveat deserves notice. I have assumed that engaging in the ambitious activity is completely tied to finding out information about oneself; that is, an unambitious agent can't observe how she would have done had she been ambitious. Although this seems reasonable for some applications [30], in the specific example I have chosen for illustration, of being in the stock market versus being in bonds, the validity of the assumption is not so clear: why shouldn't someone investing in bonds be able to pick a portfolio of stocks and see how it did?

This problem can easily be taken care of by assuming that it is costly to pick stocks well. If this cost is very high, while the returns to stocks are also very high, agents with negative, but not so negative beliefs about themselves (i.e. those who want to push their self-image back to the positive side), will only be willing to pay this cost if their money is actually at stake. Thus, they'll go to the stock market even though it's financially not worth it, because owning bonds and picking stocks to prove themselves is even worse.

1.4.5 Notes on Robustness and Psychological Accuracy

This paper started off with the premise that in a diverse array of abilities and traits, people derive pleasure from believing that they are well off, holding constant instrumental outcomes our profession generally focuses on. The psychological evidence in section 1.2 centered around facts testifying to the reasonability of this premise. However, to incorporate self-image utility into an actual decision problem and derive its implications, I have assumed a very specific structure for both the problem and the ego utility function u. While the original premise might be uncontroversial, its specific application is more problematic–we have all experienced self-image utility in one form or another, but our intuition on its functional form representation is much weaker. Since the goal of theory is not only to show what kinds of behavior are *possible*, but also to clarify how *likely* they are and *when* we should expect them, we have to worry about which results are driven by functional form assumptions on u and which ones

[30] for example, when the choice is between projects–if one has chosen to work on one project, one probably wouldn't know how the other would have turned out.

are more general.

A very special property of the model in the current section is its discreteness–the agent either chooses the ambitious or the unambitious option, and she either believes she is good enough for the ambitious option, or she doesn't. One might be worried that the two-point range of u, leading to a discrete jump in utility at zero mean beliefs, is what gives the model its mileage. It is not. The necessary ingredients of the model are twofold: the correlation of ego utility and the financial decision problem, and nonlinearities in the ego utility function with respect to beliefs.

The tight link of ego utility and the instrumental side of the problem being studied is an important tool that puts sufficient structure on the model to make it interesting. Just like any other model of individual decisionmaking, this model can't survive without *some* assumptions about the utility function. Although it is stronger than just assuming the existence of ego utility, the tie between ego utility and the instrumental problem seems to be a relatively innocuous assumption: the agent not only likes to believe that she is 'good,' but also that she is 'good enough to be investing in the stock market.' By making this minimalist structural assumption, we can capture many important facets of self-image in a tractable, strong, rational economic model that transfers easily between settings.

Non-linearities in the decision problem (and thus, due to their tight link, in ego utility) are essential to make a Bayesian model work in this setting. If ego utility was linear in beliefs about ability, the law of iterated expectations would imply that expected ego utility would never change, so that self-image would not influence behavior. Since linearity in decision rules is rather special, the non-linearity assumption is not very restrictive, either.

With these assumptions, we should see at least some of the effects discussed in this section [31]. Casually speaking, overconfidence depends on having concave parts in the ego utility function, and that some agents would stop acquiring information when

[31] One thing we lose when we move to a more general model is the ability to make clean predictions about observable behavior. It might be much easier to think about the model's implications regarding the discrete decision on whether or not to invest in the stock market than about the percentage of wealth invested.

they reach these parts: since agents expect their conclusions about their financial prowess to decrease on average when receiving information, agents here on average have too high beliefs. Self-image protection clearly also depends on having concave regions in ego utility, so just like in the original model, overconfidence is coupled with self-image protection. For example, if being more ambitious is more informative about ability, one could get overconfidence in beliefs but actions that look 'modest' relative to stated ability. Conversely, self-image enhancement happens at the convex parts of the ego utility function, and if there is sufficient pre-choice learning, self-image enhancement only arises endogenously while performing the task. Another key prediction also holds: all else equal, agents prefer information in small pieces. As noted before, this is true for *any* ego utility function, it even holds (weakly) for linear ones. Therefore, the major results of the current section carry over to more general ego utility functions as well [32].

From a psychological perspective, it is clear that there is a cognitive distance between the receipt of information and its incorporation into the psyche and/or use in specific decision problems. Along this path, any person with self-image utility has an incentive to manipulate the incoming information at least partially according to the ego's wishes. I model the cognitive distance by explicitly separating the receipt of information from its use, and the manipulability through the non-linearity. Thus, the model tries to capture the *flavor* of the incentive and ability to manipulate in a Bayesian setting, even though much of it might be non-Bayesian; non-linearities are unlikely to be the major source of manipulability. However, manipulability still has a natural meaning here: pieces of information that are more subject to interpretation are easier to 'torture' even in a Bayesian setting. I venture to guess that a non-Bayesian model equipped with a similar concept would probably produce similar results to mine, although I don't know of any non-Bayesian definition of such an idea.

[32] Two less central points I have made in section 1.4.1 are not completely general, but also depend, broadly, on the ego utility function being concave for higher beliefs and convex for lower ones. If the opposite was the case, it wouldn't be true that the more confident agents are overconfident, while the less confident are accurate. Also, it wouldn't be the case that with dispersed but on average correct priors, we would still get the overconfidence result.

Finally, this model has an even more attractive psychological interpretation. Obviously, people don't have answers to questions like 'do you expect to make money in the stock market?' ready at all times in their minds, but rather make them up when they are asked or when the question becomes relevant. The thought process through which people reach a conclusion could look very similar to the above model's sequential drawing of signals: they recall things about themselves that are relevant to the question, and can voluntarily stop the thinking process midway through. And whatever the question may be, they are motivated to find good things about themselves relative to the question. This view of the model might the most psychologically accurate one [33], although it has some theoretically unattractive features [34].

1.5 Other Applications

This section discusses other possible applications of the model developed in the paper. We move from examples where the map between the model and the applications is straight-forward toward ones where it is less so.

1.5.1 Small Businesses

The formal structure and the results of the model are readily reinterpretable for this example. It's probably harder to succeed owning a small business than in an average salaried job, and it might be desirable to be good enough to be a businessman. If this is the case, people should be in general overconfident about their ability to do well in their own business if they had one. At the same time, managing a business is more informative about ability than being employed in an average job. So, due to the self-image protection effect, we should also see talented people who seem to be

[33] Greenwald (1980), for example, also takes the view that the ego influences the memory process, comparing it to a totalitarian political regime and citing Orwell's 1984: "Who controls the past," ran the Party slogan, "controls the future: who controls the present controls the past."

[34] The interpretation introduces a slippery slope, since the question asked essentially determines the shape of the ego utility function. This is not so unreasonable when behavioral implications are concerned, but might be dangerous when communication of ability is involved: merely asking a question influences the agent's well-being.

'wasting their gifts' in trivial occupations, when they could be doing something more serious. For those already owning a small business, the self-image enhancement effect can lead them to respond too slowly to changing economic conditions: if the business does badly, they will too often convince themselves that it's not so bad, staying in a losing enterprise for too long.

A possible outcome in the current model is that the agent tries to manage a small business for a while, but, getting moderately good returns, gets out too early because of the self-image protection effect. That is, in the second period, with positive mean beliefs, she chooses the unambitious option, having a salaried job. This effect is certainly plausible in some applications: many try to write a paper or attempt to learn a sport, but drop the project after a minor disappointment. However, I think for this application the proposed effect is unlikely: getting *out* of a business is also informative about ability. One has to close books, sell real estate, and so on, all of which reflect on the quality of previous business decisions. This argument can be made more precise with the model of chapter 2, so it will be discussed further there.

1.5.2 Project Choice by Managers

It is generally acknowledged that managers have different objectives from shareholders; much of corporate finance is built around suggestions on how to bridge this difference and explanations of how current systems might already do so. One loosely articulated concern is that managers might be interested in inefficient 'empire building,' simply because they enjoy managing big companies. In my model, even if managers just derive pleasure from *thinking* they can create an empire (as opposed to enjoying the actual process of expansion or state of domination, as is assumed in other models), too many of them will tend to think they are able to do so (overconfidence). To judge whether or not this leads to too much empire building, one needs to have a sense of the relative informativeness of managing small and large enterprises. If they are approximately equally informative, too many managers will aim high and often be disappointed. In contrast, if managing a large enterprise is more informative, we won't necessarily observe overambition in the aggregate, but it won't be the right

people trying to build giants. Also, in this case the self-image enhancement motive ties empire building tightly to another form of behavior: the notion that managers take too much risk when their companies seem to be failing. When managers try ambitious projects and do badly, they might stick to it for too long, a behavior that could be called psychological gambling for resurrection.

1.5.3 Career Choice

Career choice is another relatively straight-forward application of the model. It seems clear that people attach pride to being able enough for some careers. At the same time, for many examples it is reasonable to assume that in the range of careers being considered, the ambitious one is not more informative about ability. For example, one might have to be a better actor to succeed in the movie business as opposed to in television, but one's talents will be revealed anyhow. We should then see too many young people trying to be movie actors due to an overconfidence translated into action.

1.5.4 Extrinsic vs. Intrinsic Motivation of Employees

The design of organizations could be another fruitful application of ego utility. The scarcity of simple pay-for-performance incentives and the form of existing incentives (up-or-out promotion policies, stock options, etc.) has puzzled researchers in this field (Baker, Jensen, and Murphy 1988). I will concentrate here on 'intrinsic motivation,' the idea that people might be willing to work hard for some other reason than compensation. This is mentioned extensively in the literature [35], but little insight is incorporated into formal models. In this paper's context, intrinsic motivation has a clear interpretation: when choosing what to do and how much effort to put into it, employees care not only about financial payoffs but also about how the results of the activity will reflect on themselves.

In a straightforward application of the basic model in section 1.3, workers might

[35] See Baker, Jensen, and Murphy (1988), Deci (1972), and Benabou and Tirole (1999a).

attach a certain pride to being good enough for challenging jobs. This leads them to overconfidence in what they are capable of and distortions in how they do the job. How to design incentives to take advantage of the pride factor and minimize the distortions caused by it is an interesting area for future research. For example, if the firm can hire people with a self-image protection motive (or make sure everyone acquires one quickly), it wants to create conditions in which working hard jams signals about ability.

More interestingly, some researchers argue that monetary incentives crowd out intrinsic motivation, and therefore explicit incentives might not be as desirable as standard principal-agent models would predict. Psychologists offering explanations for the experimental evidence often allude to some form of 'rationalization' effect, that people not being paid to perform a certain task justify their involvement in it by convincing themselves that it is fun to do anyway. My model is well fitted for studying rationalization of this form–the overconfidence result can be seen as agents rationalizing optimistic beliefs about themselves too often.

However, while the current model might explain intrinsic motivation by alluding to a rationalization of performing the activity, in this form it can't explain the crowding out of intrinsic motivation by monetary incentives: if in one state of the world, you want to stop collecting information about the enjoyability of the job and perform it, you also want to do it with increased incentives. To get crowding out, one needs a slightly more complicated model. Suppose the information that affects the agent's ego utility comes on two fronts: whether it's important for one's job to be enjoyable, and whether her job is enjoyable [36]. Suppose also that she prefers her job to be enjoyable only if that's important, and prefers it to be unimportant if it's not enjoyable. With no financial incentives, the agent wants to convince herself that it is important to enjoy one's job *and* that her job is enjoyable, creating an intrinsic motivation to work hard. But if she gets financial incentives to exert high effort, she'll

[36] Other interpretations of the same structure are also possible; for example, the two pieces of information might be i. whether moral considerations should enter the decision, ii. whether it is moral to perform the action. The important element is a qualifier that modifies ego utility from performance of the task.

want to exert high effort even if the job is not so enjoyable. Expecting this, she might change her information collection strategy to convince herself that enjoying one's job is not so important. But then she won't have an incentive to convince herself that the job is enjoyable, undermining the intrinsic incentive. The interesting thing about the above explanation is that it doesn't rely on a multi-period model like the multitasking explanation of the phenomenon (Holmström and Milgrom 1991), nor on implicit signals in the incentives like in Benabou and Tirole (1999a). It is simply the fact that the agent will want to respond to the incentives, *exactly what the incentives are intended to do*, that creates a self-justificatory change in world view that undermines the incentives.

Of course, it is unlikely that extrinsic motivation always undermines intrinsic motivation. A theoretical model of the above kind might delineate the circumstances in which we can expect this to be the case, a question that is not satisfactorily addressed in either the theoretical or empirical research on this topic. I intend to pursue this inquiry in future research.

1.6 Conclusion

Most of the existing information economic models treat humans as scientists objectively studying reality. According to these models, we gather information in an unbiased manner, combine it with our beliefs in a detached, logical way, reach accurate inferences and make generally good decisions. However, experimental work in social cognition has shown that we more closely resemble a scientist criticized for her work who wants to show at all costs that she is right. Partly for this reason, our inferential and decisionmaking process is full of incomplete data gathering and self-serving interpretations. The need to view ourselves favorably seems to be fundamental in motivating our behavior, and thus it should reasonably be expected to influence at least some economic decisions as well.

This paper is an attempt to model specific economically relevant consequences of our information gathering and decisionmaking behavior when this motivational side

is present. By extending the payoff space to include beliefs about the self, one can incorporate the motivational factor into models with a strictly economic methodology. I have shown that even when there is a strong restriction (Bayesian updating) on what the agents can believe about themselves, ego utility leads naturally to overconfidence in beliefs: too many people think they can do the desirable activity well relative to the number who actually can. The paper also shows that beliefs about ability are not necessarily reflected in the agent's actions: when she has good but uncertain beliefs about herself, the agent might avoid the more challenging and more informative activity precisely because she has good beliefs and wishes to preserve them. On the other hand, when she does take part in the more difficult activity and is disappointed by herself, the agent might stick with it too long in order to prove herself. Besides history- and task-dependent motivations for self-image protection and self-image enhancement along these lines, it is generally true that no one likes signals that have a great degree of potentially hope-destroying finality.

This general model of the effects of pride on behavior has a number of possible economic applications. Immediate applications include stock market participation, the decision to start a business, and project choice by managers, while a slightly more indirect connection may be made to career choice and employee motivation.

An agenda for future research would include studying the implications of ego utility in other economic applications, such as unemployment, warm glow and public goods, dangerous habits like smoking, and the intergenerational transmission of beliefs. More importantly, the model in this paper is of a rather reduced nature. Information gathering is limited to information about one's ability, which in turn affects expected financial returns in the decision periods. How the ability parameter should influence success is not explicitly modeled. By modeling the decision problem in greater detail, one can gain insight into more specific modes of self-image protection and enhancement. This is done in chapter 2.

Moving further from this paper, it would be interesting to investigate how ego utility affects strategic interactions that involve uncertainty. On the simplest level, it is clear that through overconfidence, self-image protection, and self-image enhance-

ment, ego utility can act as a commitment device for playing certain strategies with higher probabilities. This, in turn, can alter the nature of a game [37]. On a more subtle level, concern for one's self-image can influence one's view of other players' strategies, indirectly changing the game. Consider the following game theoretic model of teenage revolts. The teenager begins with a self-image enhancement motive, and can choose one of two kinds of activities: either one in which her parents can help, or one in which they can't. The parents, who want their child to succeed, then decide whether or not to help, where helping undermines the self-image enhancement motive by making success less satisfying for the child. If they don't care enough about their child's ego utility, they will want to help. Anticipating this, and out of her desire to affirm herself, the teenager 'revolts' and chooses an activity in which the parents are powerless to interfere. This outcome could be detrimental for both parties.

Even more deeply, the ego utility of the *other* player can be exploited to elicit certain forms of behavior. For example, promising to provide certain information ex post can induce her to behave differently ex ante [38].

The framework in which I'm analyzing the effects of ego utility makes at least two important implicit assumptions. First, the function u–self-image–is assumed to be externally given and fixed. This rules out the possibility that people invoke different aspects of a complex, but essentially stable, picture of the self in response to different

[37]Consider bargaining and/or 'fights': if ego utility makes it more likely that one of the players is going to fight, the other one is more likely to back down, which can be good for the first player even in the instrumental utility sense. This effect will depend on the stakes, though: for very large stakes (relative to ego utility), agents should only care about financial payoffs, so the equilibrium will look similar to one without ego utility. For very small stakes, on the other hand, the players will only want to protect their ego utility, and won't care to get the money. For medium-sized stakes, overconfidence can dominate, so that instrumentally beneficial effects of pride show through.

[38]The following is the frame for a power game between a coach and a player who has a self-image protection motive. Before the game, the player can decide to work hard or to slack. If the game goes badly, the coach observes the reasons for the outcome and decides whether or not to reveal them to the player. At the beginning, the coach might want to promise that if the player slacks off and plays badly because of this, he will 'rub it in,' making the player feel bad for making the wrong judgment. This makes slacking less attractive because of ego utility reasons. It also makes it less attractive for the player to think seriously about what she should do–she wants to give up making subjective judgments that could later reveal her to be of low ability in making them. Essentially, she might choose not to exercise her capacity to base her actions on her own judgments. The coach's ex post decision on whether or not to carry out the threat and the player's reaction to the coach's strategy can give this game a complex dynamic aspect.

threats to the person. For example, to avoid anxiety I might in general believe that the industry I'm working in is not too dangerous, but when it comes to demanding safety equipment, that's not what I think about; rather, I argue that I'm not the kind of person to work for a careless employer. Markus and Kunda (1986) have shown that such situation-dependent manipulation of an otherwise stable self-concept is possible.

This point is problematic for a model that ties the instrumental and ego uses of information so closely. However, it is also clear that that there are limits to what self-concepts we can invoke at a given time: otherwise, why wouldn't we just sit around and revel in our greatness all the time? If the informational connection between ego and instrumental utility and/or the stability of ego utility was loosened (but not completely eliminated), similar effects of a smaller magnitude would still go through.

Second, the assumption of Bayesian updating puts a very strong restriction on what the agent can believe about herself. From the study of cognitive shortcuts like the availability and representativeness heuristics, there is ample evidence that people don't update in a Bayesian way (Kahneman, Slovic, and Tversky 1982). Although there is no hard evidence that the types of distortions studied in this paper are the result of non-Bayesian updating, it seems likely that at least partly they are. This is especially true because in the model used here there is a strong sense in which the agent would *want* to be non-Bayesian: she would want to bias her signals upward at least a little bit, while still believing that they are unbiased. The ego utility gains from a small increase in the signals are first-order, while the instrumental losses are second-order. This point has resemblance in structure to one made by Landier (1999) in the context of anticipatory feelings.

Once again, there are constraints on what we can or cannot believe about ourselves– we are clearly not free to choose our beliefs or manipulate incoming information at will (Heider (1958), Tetlock and Levi (1982)). In this sense, my model is about the behavior of an agent who revels in favorable beliefs about herself, but is limited in generating them by a degree of realism. In addition, by assuming extreme rationality in information processing, I'm probably making it *more* difficult to model the behavioral effects in question. If the agent was to some degree 'allowed' to pick her

beliefs irrationally, but still had a need to know good things about herself, many of the behaviors discussed in this paper would appear even stronger. This is clear for overconfidence, but it is probably also true for self-image protection: if the agent had chosen an unrealistically high belief about the self, her incentive to avoid activities in which this could be revealed might be greater. Though not a perfect description, the Bayesian model is intended to isolate the logical consequences of utility from self-image, and many of its implications should survive in a model in which decisionmakers are less than fully rational.

Chapter 2

Ego Utility and Information Acquisition

2.1 Introduction

It is widely accepted by psychologists that healthy individuals hold unrealistically positive views of the self, and are unrealistically optimistic about the prospects they face in life. People tend to rate their own performance in tasks better than observers do and than objective criteria would warrant, to attribute success to internal causes while failure to external ones, to believe too strongly in their original judgments, and to hold excessively good views about their future prospects.

Chapter 1 argued that such an inflated self-image is most naturally understood as the result of a *desire* to hold favorable beliefs about the self. The Bayesian model of that chapter posits that agents derive pleasure, or 'ego utility,' from believing that they are capable enough to perform the more ambitious of two activities, one in which only higher types can perform well. In addition, agents' ability to control the flow of information that they receive gives them a degree of control over what they can believe about themselves. The paper shows that this is sufficient to produce overconfidence in beliefs: too many agents will honestly believe that they can perform the ambitious activity well. It also examines other logical consequences of this setup; for example, if the more ambitious of the two activities is also more informative, the overconfidence in beliefs might not be expressed in the agent's actions–although she justly believes she would make more money with the ambitious option, she chooses the unambitious one for fear of finding out otherwise.

The current chapter builds on the previous one by taking ego utility–utility derived from beliefs about the self–as given, but complicating the simplistic information structure from before. That information structure was useful for showing the overconfidence result and related phenomena, but it frames the agent's decision problem in a rather reduced form. In that setup, ability is simply taken to be a parameter that affects the agent's financial outcomes–why that parameter should influence success was not modeled. In addition, information gathering is limited to information about one's ability, whereas in the real world we gather a lot of information about the options we are about to face; one could even argue that most of our information

about the self derives indirectly from our performance in specific skill-sensitive tasks.

The first model of this chapter attempts to address both of these concerns in a model where information about ability is acquired indirectly through receiving signals about the payoffs of different financial options. Higher types receive more accurate signals, so any feedback about previous judgments is informative about the agent's ability. The decisionmaker is assumed to derive utility from her beliefs about this ability. As I argued in chapter 1, the more common consequence of ego utility is self-image protection, the tendency of people to avoid information about their ability. Therefore, most of the discussion will center around the self-image protection motive, and for simplicity, I take this motive as given and assume it in the form of the ego utility function. (One consequence of the step-function ego utility of chapter 1 is discussed in section 2.3.) I do not make an attempt to map the model in this chapter back to the previous one–such a map would preserve the effects to be discussed, but add a significant number of complications.

The model's simple setup implies that an agent with a self-image protection motive is averse to the combination of making a subjective judgment *and* reviewing it later, since the two together are informative about ability. This has three possible consequences. First, having made a subjective judgment, the agent is reluctant to review it later, since the review might put her earlier judgment in a bad light. Therefore, the agent is *sluggish* in responding to new information available to her. The flip side of this effect is that if the agent expects to have to review her judgments later, she will be reluctant to make them in the first place; in other words, she *procrastinates* in making up her mind about the choices. And finally, agents with a self-image protection motive are averse to making subjective judgments as such–they might delegate the responsibility of making assessments to a non-subjective source, even one of questionable quality. To enhance self-image, individuals might hold on to their losing choices for too long, since they are psychologically committed to thinking their earlier judgments had been right.

Section 2.4 addresses another natural extension of chapter 1's information structure. In the overconfidence model, I assume a ranking between the informativeness

of the two financial options: the ambitious option is taken to be either more, equally, or less informative than the unambitious one. For many applications, this is not the most sensible assumption. In particular, it might not be the case when only success or failure is observed in a task, not more refined information on the degree of success or failure. Then, success in a difficult task and failure in an easy one should be the most informative signals. Perhaps surprisingly, this can lead *both* self-image protectors and self-image enhancers to prefer difficult tasks, or in psychologists' terminology, to self-handicap. In a difficult task, in which most people fail, failure is basically meaningless: one just pools with most other people who also fail. On the other hand, success is really great news. Thus, even though they have exactly opposite informational preferences, both self-image protectors and enhancers are satisfied.

As in chapter 1, the models in this paper are intended to fit a number of economic applications. In section 2.5, three applications from the previous chapter, small businesses, project choice by managers, and intrinsic motivation of employees, are reconsidered in the light of the current models. For example, if intrinsic motivation is understood as a desire to manage one's self-image while performing one's job, performance bonuses might discourage workers with a self-image protection motive because they also provide information about ability. In this case, employers might deliberately want to condition pay on a noisy signal of performance and ability, thereby drowning the inferences the employee can make about herself from her pay. Depending on the worker's risk aversion, the employer might want to condition pay a lot on a very noisy signal of ability and effort, or not use incentives at all.

One particularly noteworthy application is health concerns, since it brings together effects from both chapters and points to their broader applicability. For this application, my assumption of utility from self-image is easily reinterpreted in terms of anxiety. Section 2.7 uses the methods developed in the rest of this chapter and the previous one to discuss a crucial health economics question: why people who are afraid of being diagnosed with a serious illness delay seeking professional help. Standard economic logic would say that this behavior is non-sensical: since the costs of delaying treatment are potentially much higher with a serious illness, that is exactly

the situation when people should be rushing to the doctor. I argue in section 2.7 that with anxiety the more natural outcome is *exactly the opposite*: patients will only avoid the doctor when they fear something serious. They are willing to go to the doctor for little things, but they might avoid visits with possibly very bad implications. The rest of the section considers a government health regulation problem in which the policy the government chooses also affects what people think about an underlying general health parameter that they are anxious about. In considering its policy, the government faces a commitment problem: once it has the information and it is time to implement an actual policy, and knowing how the population would react to it, the government only wants to mandate serious health measures if its benefits outweigh the ego pain inflicted on the populus. This leads the social planner to mandate the measures too rarely, even accounting for ego utility.

2.2 Modes of Self-Image Protection

Before people make important decisions that affect their life outcomes, they usually have to make judgments about the relative merits of options they are about to face. Some people are better at making these judgments than others, and it seems introspectively and observationally obvious that people like to see themselves as capable of making them well [1]. If this is the case, information is not only collected to improve decisions, but also to manage one's perception of the self. The current section models the decision-making consequences of this motivation. As in chapter 1, I will use the language of stock market participation to discuss the results, but the model is intended to be of broader applicability.

2.2.1 Basic setup

The setup of the choice problem is the following. The agent has to choose one of two options (stocks) in each of two consecutive periods. Option 1 is riskless with the

[1] For psychological evidence on this point, see the previous chapter.

return $a_1 \in (0,1)$. (This assumption is not inconsistent with the motivating example I'll be using, choosing stocks. This setup is equivalent to the agent just getting signals about the *difference* in returns of the options.) Option 2 is risky with $a_2 \in \{0,1\}$, but it pays off the same amount in both periods. Judgments are modeled as signals s^t that can be voluntarily observed about a_2 in period t before the decision has to be made. The ability to observe s^2 is not tied to having chosen option 2 in period 1. The exact timing of the problem is the following:

1: signal s^1 (choose to either observe it or not)

1': choice (payoff not observed)

2: signal s^2 (choose to either observe it or not)

2': choice (payoff not observed)

2": ego utility realized

3: financial outcomes realized

As an example, the following choice problem has a time structure resembling the above. The agent gets an opportunity to learn about and invest in a firm which will eventually succeed or fail. Later, when new information about the firm is available, the agent can once again decide whether to invest. The actual financial outcome is further down the line. Choosing to observe the signal s^1 or s^2 corresponds to making a judgment or reviewing a judgment about the firm (stock 2), respectively [2]. After possibly reviewing the options and choosing one of them, the agent has to confront how what she has seen reflects on her ability; then, her utility from self-image (ego utility) is realized. Since in the present chapter we are primarily concerned with

[2]Modeling judgments as signals is a simplification. A more realistic view is that the agent collects decision-relevant information, and the real judgment she has to make involves deciphering this information, for example by choosing relative weights of importance. It is probably the latter step that better investors can make better. In my formulation below, the mental process of making a judgment is collapsed into a reduced form.

information gathering for choice, we assume that the financial outcomes are realized so late as not to affect the ego; if the information implicit in the observation of financial payoffs also affected the ego, the discussion would be cluttered by many additional cases and effects.

The signal s^t is imperfectly correlated with the actual payoff of option 2. In particular, the space of the signals is also $\{0,1\}$, and the probability that one is 'right' varies with the agent's type and the nature of the signal. We distinguish between two kinds of signals. The probability that a *type-dependent* signal in period t is right can be one of two values:

$$Prob(s^t = a_2) = q^t = \begin{cases} q_H^t & \text{if agent is high-type;} \\ q_L^t & \text{if agent is low-type.} \end{cases} \tag{2.1}$$

We assume $q_H^t > q_L^t \geq \frac{1}{2}$. First-period signals are always type-dependent to capture the notion that early decisions, when things are usually not so clear yet, depend more on a subjective judgment. In contrast, for the second period, we will consider both the case of type-dependent and type-independent signals. The latter kind of signal is accurate with probability $q_I > \frac{1}{2}$, independently of the agent's type. The agent's priors are summarized in $p_{jk}^0 = Prob(a_2 = j, q^t = q_k^t)$, and I use the notation $p_{jk}(S)$ for the agent's posteriors after observing the set of signals S. Let S^t the set of signals observed by the end of period t.

Utility from self-image in this problem depends on the (Bayesian) agent's subjective probability of being a high type. At the end of period 2, this probability is given by $p_{1H}(S^2) + p_{0H}(S^2)$. Total utility is then

$$wu\left(p_{1H}(S^2) + p_{0H}(S^2)\right) + n_1 a_1 + (2 - n_1)a_2, \tag{2.2}$$

where n_1 is the number of times the agent chooses option 1. The agent is a Bayesian expected utility maximizer; see chapter 1 for a justification of this kind of model.

$w > 0$ is simply a weighting parameter. As noted before, this section focuses on

self-image protection, so we assume that u is strictly concave [3]. There are two reasons to do this. First, chapter 1 argued that self-image protection is likely to be the more common phenomenon affecting agents with ego utility. In addition, in a model of this type, self-image protection is more interesting: it counterbalances the classical value of information.

2.2.2 Preliminary results

We are primarily interested in what kinds of signals or combinations of signals are informative about the agent's type, since this is what the interesting effects will depend on.

We start with two obvious facts.

Fact 4 *Any combination of type-independent signals is uninformative about ability.*

Fact 5 *(Risk neutrality) In each period $t = 1, 2$, the agent chooses option 1 if and only if*

$$a_1 > p_{1H}(S^t) + p_{1L}(S^t), \tag{2.3}$$

That is, the agent maximizes the expected return conditional on her information: she chooses option 1 if and only its return is higher than the posterior probability that option 2 would yield an outcome of 1.

The following lemma is a key intermediate result. Its proof is relatively straightforward, but requires a few steps, so it is relegated to the appendix.

Lemma 1 *A type-dependent signal combined with any other signal is always informative about ability.*

Although it is not quite accurate in general, the intuition for the case when $p^0_{0H} = p^0_{1H}, p^0_{0L} = p^0_{1L}$ is the most useful to understand. In that case, a type-dependent

[3] Analogously to risk aversion, this means that the agent is 'information averse' with regard to her ability–her expected utility decreases with the addition of information.

signal by itself is not informative–since both outcomes for option 2 are equally likely a priori, making a judgment either way doesn't say anything about the person. Then, since high types are more likely to receive 'correct' signals, receiving consistent informative signals, or, in plainer terms, 'not getting confused,' is a sign of being a good decisionmaker. And since any two signals are either consistent or inconsistent, the two signals are informative [4].

2.2.3 Sluggishness and Procrastination

We concentrate on the case of independent and neutral priors. Let $r_2 = Prob(a_2 = 1) = \frac{1}{2}$, and assume that the prior probability c of being a high type, the trait relevant for ego utility, is independent of a_2. This probability will be interpreted as confidence; thus the notation c. We assume that $0 < c < 1$, so that the agent is not completely certain of her ability. Note again that due to the timing of the problem, observing the financial outcome does not convey information that enters ego utility.

The next two theorems constitute the main results of section 2.2. They show, respectively, that if ego utility is important enough for the agent, then she will either fail to reconsider her choices, leading to a sluggishness in them, fail to make up her own mind about it at the first given opportunity (theorem 7), or, in certain conditions, wait for a type-independent signal to make up her mind (theorem 8).

Theorem 7 Sluggishness and Procrastination *Suppose that s^1 is type-dependent. If her ego utility is sufficiently important (w is sufficiently large), the agent will observe exactly one of the signals.*

Proof. That the agent won't observe both signals for a sufficiently large w is an obvious consequence of lemma 1. To show that one signal will in fact be observed, we show that a single signal is not informative about ability. For type-independent

[4]Using the proof of lemma 1, it is easy to show that *if* the type-dependent signal is uninformative by itself, then the good news about ability is if the two signals are the same. So in that case the intuition is still correct. It is not correct in general, though: if, for example, the agent is sure that $a_2 = 1$ and both signals are type-dependent, then it is better to receive a zero signal and a one signal than receiving two zero signals.

signals, this is implied by fact 1. For type-dependent signals, it follows from neutral priors: since $a_2 = 0$ and $a_2 = 1$ are equally likely, both types of agents receive the signal $s^1 = 1$ with probability $\frac{1}{2}$. Formally, we have

$$Prob(q^t = q_H^t | s^t = 1) = \frac{\frac{1}{2}cq_H^t + \frac{1}{2}c(1-q_H^t)}{\frac{1}{2}cq_H^t + \frac{1}{2}c(1-q_H^t) + \frac{1}{2}(1-c)q_L^t + \frac{1}{2}(1-c)(1-q_L^t)} = c. \tag{2.4}$$

This completes the proof. □

The following corollary illustrates the use of this theorem for $a_1 > \frac{1}{2}$, a case when the agent's prior beliefs favor option 1.

Corollary 1 *Suppose that $a_1 > \frac{1}{2}$ and that the first-period signal is type-dependent and the second one is type-independent. If ego utility is sufficiently important (w is sufficiently large), only one of the signals will be observed. It will be the second one if and only if*

$$2(cq_H^1 + (1-c)q_L^1 - a_1)_+ < (q_I - a_1)_+. \tag{2.5}$$

Proof. We know from theorem 7 that for a large enough w exactly one of the signals will be observed.

If only one of the signals is observed, it will be the one with greater instrumental value. A signal has instrumental value if it can reverse a decision; and its value is the probability of reversing a decision times the difference in conditional expected utilities. With $a_1 > \frac{1}{2}$, a decision can only be reversed when the signal is favorable. Now both kinds of signals will equal 1 with probability $\frac{1}{2}$. It is also easy to show that

$$\begin{aligned} Prob(a_2 = 1 | s^1 = 1) &= cq_H^1 + (1-c)q_L^1 \\ Prob(a_2 = 1 | s^2 = 1) &= q_I. \end{aligned} \tag{2.6}$$

By fact 5, the decision will be reversed after $s^1 = 1$ if $cq_H^1 + (1-c)q_L^1 > a_1$ and after $s^2 = 1$ if $q_I > a_1$. The respective differences in expectations are therefore $(cq_H^1 + (1-c)q_L^1 - a_1)_+$ and $(q_I - a_1)_+$. Furthermore, since the first period's signal is

there to affect both periods' choices, it will be preferred if it is at least half as valuable as the second. These conditions are summarized in inequality 2.5. □

It's easy to generate decision rules for the other cases of the problem, but those are not worth writing down for our purposes.

This theorem summarizes two basic behavioral distortions that can arise as a consequence of ego utility. The first one I have labeled sluggishness: once the agent has made a judgment whose accuracy depends on her ability, she will be reluctant to look at new information later, afraid the new information would reveal the judgments she has made to be poor. Without new information, of course, the agent will choose the same option as before (see fact 5), exhibiting an excess sluggishness in changing options relative to the information available to her.

The second distortion to financial decisionmaking arises when the second-period signal is more instrumentally valuable than the first-period one. When the agent knows her choices will be evaluated in the future, she might not want to think about them seriously today, so that the later judgments are not reflexive of her ability. That is, she puts off (procrastinates) making a serious decision.

In the above setup, whether the information aversion of the agent played itself out in sluggishness or procrastination depended only on the informativeness of s^1 and s^2. But there is more to it than that. By fact 4, type-independent signals observed in isolation (not in combination with type-dependent signals) are not threatening to the ego, so agents have a general preference for type-independent signals irrespective of informativeness. Specifically, agents who expect to receive feedback about their choices later or to possibly reconsider them should have an incentive to wait for a type-independent signal even if it is not more accurate. To make this statement formally, one needs to expand the basic model a little bit. This is done in theorem 8.

Theorem 8 Deferral to objective criteria *Consider the same setup as in section 2.2.1 with the following modifications. Now there are T periods of choice, with the same timing pattern as periods 1 and 2 in section 2.2.1, and the non-ego utilities are realized in period $T+1$. Suppose s^1 is type-dependent.*

1. *If s^2 is type-independent, and there is a type-independent signal s^3 that has to be observed by the agent, then for a sufficiently large w only s^2, not s^1, will be observed.*

2. *If $s^2, s^3, \ldots, s^T$ are type-independent and s^1 is not perfectly informative ($c = q_H^1 = 1$), then for a sufficiently large w and a sufficiently large T and sufficiently informative $s^3, \ldots, s^T$ signals, s^2, and not s^1, will be observed.*

Proof. Immediate from fact 4 and lemma 1. □

There are other variants of the same principle. For example, the above theorem still holds true if s^2 is type-dependent, but does not distinguish the types well (for example, when q_H^2 and q_L^2 are close). Or, even with two periods, if the second-period signal is type-independent and in the first period, the agent can choose between a type-dependent and a type-independent signal, then for a sufficiently large w she will prefer the type-independent one even if it is less informative [5].

Procrastination in theorem 7 and theorem 8, then, describe different aspects of the unwillingness of agents to 'make up their minds' about choices that will have to be reviewed for some reason. One can defer this responsibility by relying on previous knowledge or tradition (priors) or by recruiting help that is not reflective of the self.

In section 1.3 of chapter 1, we saw the possibility that people don't enter the stock market when it seems financially sensible to do so because they are averse to finding out they are bad investors. Now, on top of that financial distortion, there are other possible effects that complicate the agent's behavior even if she enters the stock market. For one, she might not put too much effort into it, so that the quality of her judgments is not reflective of her ability (procrastination); or, having made a judgment, she might not want information on how her stocks are doing (sluggishness). In addition, people might try to make their choices by scrutinizing questionable 'objective' information instead of accepting that the decision might be inherently sub-

[5] Consider a group of hikers coming to a fork in the road, having little idea which way to go. If they can't figure out the likely way, many people in this situation prefer to flip a coin, even though that can't possibly increase the probability of making a good choice relative to a subjective judgment.

jective. Again, in personal finance it's not clear if financial advisors are worth the money other than for lifting responsibility off one's shoulders.

2.2.4 Confidence, Feedback, and Behavioral Distortions

Having reviewed the basic behavioral distortions motivated by self-image protection, it is natural to ask how these effects respond to different environments. We will consider two issues. First, motivated by the paper's general focus on self-image, we examine how the agent's prior probability of being a high type (c) influences the two effects. This parameter is interpreted as confidence, and the posterior probability of being a high type is what ultimately determines ego utility. Second, we look at the consequences of giving the agent feedback about the stocks available to her. This is interesting because it is the economist's remedy for limited information gathering.

In the setup of corollary 1, the higher is the agent's prior probability of being a high type, the more likely it is that condition 2.5 is violated; that is, if it's violated for some c, it is also violated for $c' > c$. Thus we have the following.

Theorem 9 *Suppose s^1 is type-dependent and s^2 is type-independent. If the agent is sluggish for some $0 < c < 1$, she does not procrastinate for any $c' > c$.*

This theorem says that in some sense confidence helps in overcoming the procrastination problem. It is a version of the effect found in a static setting by Weinberg (1999), where higher signals about ability lead the agent to take on more challenging tasks, which she would otherwise avoid.

The effect in theorem 9 is clearly driven by the fact that if c is higher, the first-period signal is perceived to be more informative by the agent, and so harder to give up. This intuition goes quite far. A first-period signal, at the very least, improves the first-period decision, and an agent who sees herself as a better decisionmaker will think it improves that decision more. Although there might be complications compared to the simple setup of theorem 9, this intuition is not reversed [6]. The

[6]For example, if s^2 is also type-dependent, then a higher c also makes the second-period signal

only crucial caveat is that agents who are more certain of their ability will suffer less from *both* procrastination and sluggishness: they can learn less about themselves through making subjective judgments. In this setup, agents with confidence levels close enough to 0 or 1 won't avoid any signals. Again, an effect similar to this was also found in Weinberg (1999). So, to make an earlier statement more accurate, confidence helps in overcoming procrastination as long as it doesn't at the same time make the agent more unsure about her type.

But in a multi-period problem there is a flip side to the beneficial effects of confidence against procrastination. In contrast to what I have argued above for signals in the first period, it seems that signals in the second period might be less likely to be observed by confident agents. Consider, for example, $a_1 > \frac{1}{2}$, $s^1 = 1$, and s^2 type-dependent with equivalent informativeness to s^1; that is, $q_H^1 = q_H^2 = q_H$ and $q_L^1 = q_L^2 = q_L$. Further, make the assumption that $Prob(a_2 = 1|s^1 = 1) = cq_H + (1-c)q_L > a_1$. The informativeness of the signal s^2 stems from the possibility that $s^2 = 0$, which leads to a reversal of the decision for any c: $Prob(a_2 = 1|s^1 = 1, s^2 = 0) = \frac{1}{2}$ for any c. So conditional on reversing the decision, the value of doing so is the same for all c. However, the probability (or the perceived probability) that the decision will be reversed is *decreasing* in c:

$$Prob(s^2 = 1|s^1 = 1) = c((1-q_H)^2 + q_H^2) + (1-c)((1-q_L)^2 + q_L^2), \tag{2.7}$$

which is increasing in c. Confident agents think that they can make good decisions the first time around, so they think it is less likely they would change their minds. Therefore, they don't care about reconsidering the decision too much.

This highlights a key distinction between early and late signals: while more con-

more informative. However, for this to outweigh the effect coming through the accuracy of the first-period signal, it has to be very strong (since the other signal is observed earlier). In particular, as long as $2(q_H^1 - q_L^1) > q_H^2 - q_L^2$, theorem 9 still holds true, with the proof being the same. And I would expect this relationship to hold in general: it is likely to require less skill to make a judgment later rather than earlier. (It might be the case that information comes in between periods 1 and 2 that only high types can analyze, making $q_H^2 - q_L^2$ possibly higher than $q_H^1 - q_L^1$. However, in that case, the second-period decision is likely to be quite hard–the signal not being so informative,–so the first-period signal is observed for any c.)

fident agents always consider a first-period signal to be more informative than less confident agents, they might find a later signal less informative if they have already observed a signal earlier. This is the case when the crucial question is whether to reverse decisions. Thus, while confident people are likely to be less prone to procrastination, they are probably more prone to sluggishness.

Let us move on to the solution most commonly recommended by economists against ignorance: feedback, or, more generally, information. Intuitively, the problem is that agents sometimes don't want information because they want to protect their egos. Feedback, by disallowing them from getting caught up in an ignorance state, might be a remedy. Of course, the decisionmaker can in general avoid performance-relevant information at least for a while, no matter how hard it is forced on her [7]. Still, if our agent knows that she won't be able to 'flatter herself' in the long run, she might as well not flatter herself now. The logic is the same as in the case of an academic who likes to think her paper is really good and therefore tries not thinking about it too much, but if she has submitted it for publication and knows there is a judgment coming up about it soon, she will be more realistic. This intuition, related to the well-known fact in the psychology literature that self-serving biases diminish with the threat of verification (Fiske and Taylor 1991), is the core of the following very general theorem.

Theorem 10 *Take a T-period model as in theorem 8. Suppose some period i's signal s^i becomes available again in period $j > i$; that is, $s^j = s^i$. Then if the agent observes s^i if s^j doesn't have to be observed, she observes s^i when s^j has to be observed.*

Proof. In any state of nature in period i, consider the expected utility from following the optimal policy after observing vs. not observing s^i. The expected utility when observing s^i is independent of whether s^j has to be observed, while not

[7] It is an interesting, and clearly crucial, question on how one can provide feedback to someone who doesn't want to hear it. I'm not going to deal with this in any detail here.

observing it yields a (weakly) lower expected utility if s^j has to be observed. Thus if s^i is observed when s^j doesn't have to be observed, it is also observed when s^j has to be observed. □

The converse is clearly not true, so forced feedback can undermine ignorance. This proof takes strong advantage of the fact that s^i and s^j are the same signal. How realistic is this assumption? There is at least one sense in which it is not: s^i could be a type-dependent signal, and it seems unreasonable to assume that anyone who might be providing the feedback would have access to that signal at any time. If the signals are not the same, and especially if s^i is type-dependent and s^j is type-independent, the above theorem does not hold in any generality. Consider the following example:

Lemma 2 *Consider the two-period model with s^1 type-dependent and s^2 type-independent. If s^1 is observed when s^2 has to be observed, it is also observed when s^2 doesn't have to be observed.*

Proof. When the agent has to observe s^2, the value of observing s^1 decreases weakly both in the instrumental and ego-utility senses. □

When there is an unavoidable 'reality check' next period, we tend to hesitate more in making up our opinions today–being more afraid that our inferences will turn out wrong. Therefore, it seems that environments with good feedback can help make sure agents psychologically committed to believing in a judgment look at objective information carefully when reviewing that judgment, but promising that very feedback can undermine their willingness make an assessment in the first place. Feedback provides a good learning environment for mistakes, but it is an environment that we might not want to enter.

2.3 Self-Image Enhancement and a Sunk Cost Fallacy

As section 2.2 demonstrated, self-image protection has a variety of manifestations in information acquisition about and the choice between stocks. Self-image enhancement

is perhaps less interesting in this setting: since both ego and instrumental considerations compel the agent to accept free information, there won't be any (apparent) distortions in behavior [8]. Thus it is not worth solving the model of section 2.2 for a convex u. Instead, to show one interesting phenomenon involving self-image enhancement, we go back to the step-function ego utility of chapter 1. Since under reasonable assumptions the enhancement motive arises only in the second period after a bad outcome in the first, we assume that the decisionmaker has entered the stock market, made a subjective judgment in favor of one of the stocks, invested in it, and suffered a loss. She now has negative mean beliefs about herself, and must decide whether or not to hold the stock. We assume that the distribution of excess returns of the currently held stock (relative to the other stock) is centered around zero, and the agent can observe three identically distributed type-independent signals about it [9]. We keep the assumption from section 2.2 that financial payoffs are realized after ego utility. The characteristic financial behavior that results is a form of an apparent *sunk cost fallacy*: since the investor is psychologically committed to believing she has made the right call, she will tend to hold on to her losers too much. In the aggregate, those who sell do better in the market than those who hold. From this pattern, one might otherwise be tempted to conclude that the agents are acting irrationally, but here there is no real sunk cost fallacy going on–every participant does the financially sensible thing based on her information!

It's easy to see why this happens. By receiving a positive signal about the currently held stock's relative return, the agent's past judgment will be put in a better light again, and this might put her back up on the positive side of ego utility. Thus, it is possible that she will stop acquiring information when she has seen a positive signal.

[8] One could of course build a model in which self-image enhancement leads to costly information acquisition about stocks. I'm not going to do this here.

[9] These are just simplifying assumptions that allow us to abstract away from price determination in the market. For the result that agents hold losers too often, all we need is that news about the investment that would put beliefs on the positive side again are better than news that would make the agent hold. Assuming that without any new information the agent would want to hold, this is always the case. Theorem 11 wouldn't be meaningful without the symmetry assumption. And I'm using three signals so that after one signal, there could still be a reversal in which stock the agent chooses.

In contrast, she will not stop acquiring information if she has seen a negative signal, since she has nothing to lose on the ego utility side. In addition, as in section 2.2, all agents do the financially sensible thing conditional on what they know: they choose the stock with the higher expected return. This is the consequence of our assumption that financial outcomes are realized after ego utility. Therefore, those who end up selling trade on better information, and so on average will do better.

Theorem 11 *Consider an agent with negative beliefs about herself, holding a stock whose excess return relative to the other stock is distributed symmetrically around zero, and whose return is positively correlated with type. If the agent can observe three i.i.d. type-independent signals about the stock's relative return, the expected relative performance of her choice of stock conditional on selling is greater than the expected relative performance of her choice of stock conditional on holding.*

Proof. If the agent doesn't stop acquiring information after one signal, the two expectations are clearly the same. If she does, those people who hold all have one signal to support this choice, while those who sell have one or three. Therefore, the expected profit of those who sell is greater. □

Consider again our statistician observing a population of these agents. Suppose she looks at the performance of the subgroup who sell their losers and compares these to the subgroup of agents who hold. She will find that on average those who sell do better. This is the sort of comparison Odean (1998) makes to claim that small investors are making bad financial decisions, although he compares the performance of winners sold to the performance of losers held. In this model, the extent to which agents' decisions are financially wrong is limited to information gathering. Since everyone picks the right stock conditional on their information, it's not true that any of the holders should switch their strategy and instead become sellers. Agents are simply psychologically committed to *thinking* that their old judgment was good, and so on average will rationally believe that they were right more often than they actually were, achieving this through being *uninformed* in an asymmetric way. It is the uninformedness that leads to the above aggregate pattern in returns, and this can

happen in other models with private information as well.

As a final note on the last two sections, we remark that, somewhat surprisingly, self-image protection and enhancement often lead to similar observed behaviors: agents hold on to their assets too long. For self-image protection, this happens because the agent is unwilling to consider new information, and with self-image enhancement, because she wants information that, along with validating her past judgment, also justifies holding on to her investments [10]. This sets the stage for the next model, in which both kinds of agents prefer similar kinds of signals, not only behave similarly.

2.4 Self-Handicapping

This section demonstrates a general result on preferences for information of agents with ego utility. Whereas section 2.2 focused on self-image protection and its effects when multiple signals are observed, here I focus on a single signal or task and don't restrict the ego utility function to be concave. Perhaps surprisingly, this does not necessarily prevent us from making statements about what kind of signals the agent prefers. Even though self-image protectors and self-image enhancers have exactly opposite informational preferences, in general they only rank informative tasks or signals in the opposite way if the tasks are rankable in informativeness. In many real-world tasks, only success or failure is observed, not more refined information on the degree of success or failure. In that case, success in a difficult task and failure in an easy task are the most informative signals, and thus the tasks are not rankable in informativeness. For a very difficult task, in which most people fail, but a few very good ones succeed, failure is basically meaningless: one just pools with most other people who also fail. Self-image protectors like this. On the other hand, success is really great news, which self-image enhancers like. Whether the agent is in for one or the other, the task might be ideal for her. The current section formalizes this

[10] Theoretically, it is also possible that the agent *sells* a loser too easily. This can happen if her mean beliefs haven't dropped below zero, and the beliefs that would make her hold are higher than the beliefs that would keep her satisfied about her ability. However, with mean beliefs above zero, the distortion could go either way: one could get too much or too little selling of losers.

intuition.

Since the point is independent of instrumental considerations, I use a model with only ego utility. There is a single task or signal to choose. For each available task, there are two possible observed outcomes: success or failure [11]. The agent has an underlying ability parameter q, and for notational simplicity the tasks are parameterized by the cutoff ability level q^c required to succeed in them. Therefore there is no risk in the performance outcome other than that arising from an imperfect knowledge of ability; the results would be identical if a signal was generated based on the underlying q and the tasks were parameterized by the cutoff level for this signal. We assume that u is increasing and twice differentiable in the expectation of q. We will consider both information-averse (concave) and information-seeking (convex) ego utility functions u, where I'm using these terms analogously to risk-aversion and risk-seeking with standard utility functions: since information decreases expected ego utility for a concave u [12], agents with such a utility function will be averse to information, and conversely for a convex u. We assume that there are a continuum of tasks, $q^c \in (\underline{q}, \bar{q})$. The differentiable probability distribution function $f(q)$ summarizes the agent's priors about ability on the interval $(\underline{q}, \bar{q})$. We are interested in the agent's preferences over the cutoff value $q^c \in (\underline{q}, \bar{q})$.

The expected ego utility of the agent when participating in the task with threshold ability level q^c can be written as

$$Prob(q < q^c)u(E(q|q < q^c)) + Prob(q > q^c)u(E(q|q > q^c)). \tag{2.8}$$

Differentiating this with respect to q^c gives $f(q^c)$ times [13]

$$u'(E(q|q < q^c))(q^c - E(q|q < q^c)) + u'(E(q|q > q^c))(E(q|q > q^c) - q^c)$$

[11] Tying this back to the model of chapter 1, this is analogous to observing the payoff +1 or -1 when the ambitious option is chosen in period t, instead of observing s^t.

[12] By the law of iterated expectations, the expectation of q is unchanged with new information. Then, Jensen's inequality implies that expected utility decreases for a concave u.

[13] We have

$$E(q|q > q^c) = \frac{\int_{q^c}^{\bar{q}} qf(q)dq}{\int_{q^c}^{\bar{q}} f(q)dq}, \tag{2.9}$$

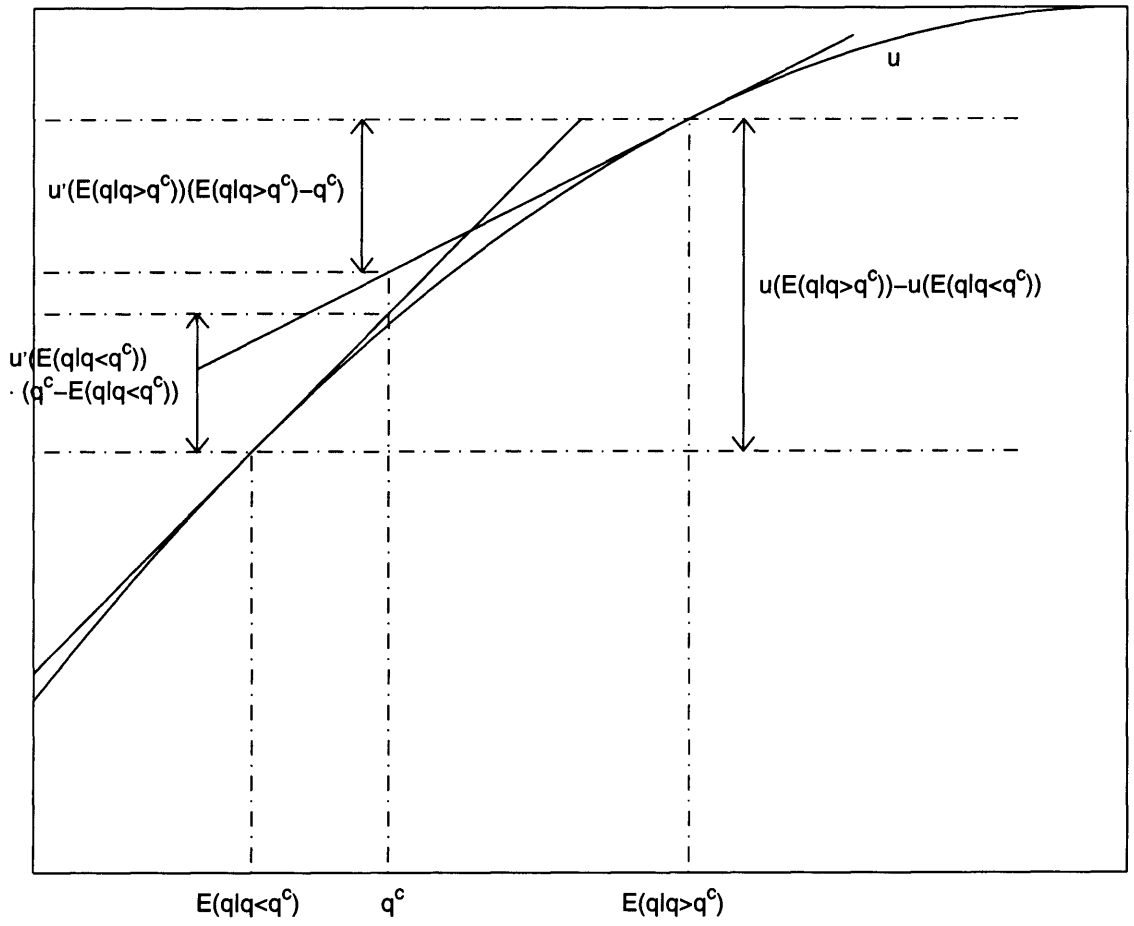

Figure 2-1: A graphical representation of the derivative of the agent's expected ego utility with respect to q^c

$$- \quad [u(E(q|q > q^c)) - u(E(q|q < q^c))]. \tag{2.11}$$

The second derivative is $f(q^c)$ times

$$\begin{aligned} & u''(E(q|q < q^c))(q^c - E(q|q < q^c))^2 \frac{f(q^c)}{Prob(q < q^c)} \\ + \quad & u''(E(q|q > q^c))(E(q|q > q^c) - q^c)^2 \frac{f(q^c)}{Prob(q > q^c)} \\ + \quad & u'(E(q|q < q^c)) - u'(E(q|q > q^c)) + \frac{f'(q^c)}{f(q^c)^2}(\text{first derivative}), \end{aligned} \tag{2.12}$$

where it's not worth writing down the first derivative again.

We will show that the derivative 2.11 is everywhere positive for both sufficiently information-averse and sufficiently information-seeking ego utility functions u [14]. Note that if u is information-averse enough ($\frac{u''}{u'}$ is sufficiently negative everywhere), then the second derivative 2.12 is negative, while for sufficiently information-seeking u it is positive. Given the sign of the second derivative, to complete the proof that the derivative 2.11 is positive everywhere, it is sufficient to show that for a concave u, the derivative is positive for q^c close enough to $\bar{q}$ and for a convex u, it is positive for q^c close enough to $\underline{q}$.

This is easy to see from a graphical interpretation of the derivative, which is illustrated in figure 2-1 for a concave ego utility function u. The tangent to u at the point $E(q|q < q^c)$ has slope $u'(E(q|q < q^c))$, so $u'(E(q|q < q^c))(q^c - E(q|q < q^c))$ is equal to the difference in values of the tangent line evaluated at points q^c and $E(q|q < q^c)$, as shown. Similarly for $u'(E(q|q > q^c))(E(q|q < q^c) - q^c)$. It's clear that the derivative is greater than zero if and only if q^c is to the right of the intersection point of the two tangents. Now if q^c is close to $\bar{q}$, we have $E(q|q > q^c) \approx q^c$, so this

---

whose derivative is

$$\frac{-\left(\int_{q^c}^{\bar{q}} f(q)dq\right) q^c f(q^c) + \left(\int_{q^c}^{\bar{q}} qf(q)dq\right) f(q^c)}{\left(\int_{q^c}^{\bar{q}} f(q)dq\right)^2} = \frac{f(q^c)}{Prob(q > q^c)}(E(q|q > q^c) - q^c). \tag{2.10}$$

Similar manipulations give the rest.

[14]The results can also be understood as local statements for sufficiently information-averse or information-seeking *parts* of u.

is the case. Symmetric reasoning shows that for a convex u, the derivative is positive if and only if q^c is to the left of the intersection point, which is satisfied when q^c is close to $\underline{q}$.

In essence, we have shown that agents with either a self-image protection or a self-image enhancement motive prefer to self-handicap, that is, to choose or even create excessively difficult tasks. Moreover, these utility functions, even though one should lead to information seeking and the other to information avoidance, rank the given informative tasks in exactly the same way! In other words, people prefer to see signals about their ability in which they can only win–either they find out good news or practically no news at all. Although the extreme preference for self-handicapping (that agents prefer q^c to be as large as possible) is only true for sufficiently strong information preferences, the intuition behind the result is quite robust. For self-image protection, one wants to avoid bad downside news, so pooling with a lot of failures is good. For self-image enhancement, one wants good upside news, so success should be informative. Self-handicapping takes care of both.

Conversely, for both very information-averse and very information-seeking agents the least preferred kinds of news about one's ability are those that can only turn out badly. When basically everyone succeeds, success won't make us proud, while failure will make us excessively sad.

2.5 Other Applications

2.5.1 Small Businesses

It is reasonable to assume that people, and especially those contemplating starting a business, derive ego utility from thinking that they are good enough to do so. As discussed in chapter 1, if this is the case people should be in general overconfident about their ability to do well in their own business if they had one, but if managing a business is more informative about business sense than being employed in a salaried job, some people might stay out because of self-image protection.

The model of section 2.2 applies to those already owning a small business. Due to sluggishness, they will try to avoid new information on their decisions as long as they can, leading them to respond too slowly to changing economic conditions. Theoretically, it is also possible that after having made a judgment about the business, the agent gets out too early to avoid receiving further signals about the quality of her judgment. However, this is only the case when getting out is less informative about previous judgments than staying in. The tasks associated with getting out, of course, also convey a lot of information about the quality of previous judgments; and, at the very least, this information could be delayed if one goes along casually managing the enterprise. Thus, although the agent might not *enter* into a business activity because of the self-image protection motive, this does not mean that she will *exit* because of it. What changes fundamentally when one enters is that a judgment is made about what to do, breaking the ex ante relation between self-image protection and being out.

In addition, if the business does badly, according to section 2.3, agents will too often convince themselves that it's not so bad, staying in a losing enterprise for too long. Theorem 11 in section 2.3 also predicts that under some conditions, those who close up shop after suffering losses do better financially than those who continue, a result of the psychological commitment to earlier judgments about the business.

2.5.2 Project Choice by Managers

We have seen in chapter 1 that an empire-building ambition, understood as ego utility from the belief that one can manage big enterprises, can result in distorted project choices by managers. But whether or not project choice itself is distorted by the manager's self-image protection or enhancement motives, these should distort *how* she manages the chosen project. In particular, she is likely to be prone to all the distortions in instrumental decisions outlined in section 2.2. On the self-image protection side, managers might want to hold off on decisions until it's clearer what to do (procrastination), but once they have made an important judgment call, they'll be reluctant to reexamine it.

If owners don't want their managers to fall prey to the above weaknesses, what can they do [15]? If the most important thing is to avoid procrastination, the results in section 2.2 suggest that hiring confident (even overconfident) managers might be a good idea. These people come in with enough conviction to affect the quick changes that might be necessary in the marketplace. A clear downside is that the changes could be excessive, at least if the manager is overconfident. At the same time, confidence makes the sluggishness problem worse, so the drastic changes could stick around for a long time [16]. Of course, there are things not modeled in the present paper that the owners might do to alleviate this problem; for example, they might want to switch management often, bringing in 'fresh blood' in the form of people whose egos are not threatened by a review of earlier judgments.

2.5.3 Extrinsic vs. Intrinsic Motivation of Employees

Pay-for-performance systems necessarily involve distinguishing workers based on their performance, and there is a well-articulated notion that this might make those with a smaller bonus feel bad and eventually erode employee morale. In a model with ego utility, this has a natural meaning: the amount of compensation a worker gets is not only a signal about her performance, but indirectly also about her ability. Workers with a self-image protection motive don't like this and so have to be compensated for the 'ego pain' inflicted on them by the system. In addition, they want to take actions that make the bonus less informative about ability.

This section examines one trick employers might do if they still want to provide incentives: they might purposefully use noisy signals to condition pay on, thereby drowning the inferences the employee can make about herself from her pay.

Formally, we build on the model of section 2.2. The agent can observe a type-dependent signal about project payoffs, with higher types getting more accurate signals. However, we make a few changes to make the point more easily. We assume

[15] Standard performance-based financial incentives are discussed in section 1.5.4.

[16] Fortune magazine's June 21, 1999 lead story on 'Why CEOs Fail' essentially identifies sluggishness as the number one CEO killer. Denial, as the magazine calls it, seems to be worst for very subjective decisions: the company's business model or subordinates selected for key positions.

there is only one period, but the amount of final compensation is revealed before the agent's ego utility is realized, so that it also provides information about ability [17]. To avoid the trivial solution that no incentives are necessary, we also assume that there is a utility cost $\epsilon > 0$ of observing the signal. And finally, the agent is now risk-averse: her utility function for monetary outcomes is v, where v is concave. Other than these changes, we operate in the neutral, independent setup of section 2.2.3.

In addition, we have to define the employer's problem. I assume that the employer observes the outcome a_i if the agent chooses project i, and can condition pay on this outcome. Let the conditional wages be w_0, w_1, and w_{a_1}. However, the owners of the firm can introduce some noise into the amount of compensation; in particular, if the agent chooses option 2, they can mix the actual outcome with a purely random signal, paying w_1 with probability $p_H = tq_H + (1-t)q_L$ and w_0 with probability $1 - p_H$ [18]. I denote by α the probability that the actual outcome is used for pay.

We assume that the employer wants to give the agent incentives to observe the type-dependent signal that is available to her, and choose option 2 if and only if that signal is good. Thus, we don't study the employer's full problem, only the implementation of one kind of agent behavior.

We start with the agent's updating problem. For $w_0 \neq w_1$, let

$$\begin{aligned} d_H(\alpha) &= Prob(q = q_H | s_1 = 1, wage = w_1) = \frac{tq_H\alpha + (1-\alpha)p_H}{p_H} \\ d_L(\alpha) &= Prob(q = q_H | s_1 = 1, wage = w_0) = \frac{t(1-q_H)\alpha + (1-\alpha)(1-p_H)}{1-p_H} \end{aligned} \tag{2.13}$$

$d_H(\alpha)$ and $d_L(\alpha)$ are the agent's posteriors about her ability when she observes a positive signal about project 2, chooses it, and then receives wages of w_1 and w_0, respectively. For example, when $\alpha = 0$ (when the payoff is based on a purely random signal,) $d_H(\alpha) = d_L(\alpha) = t$–the agent's payoffs are not informative about her ability.

[17] This makes the updating problem very similar to the two-period model, because there are two signals altogether (the type-dependent followed by the amount of compensation), which is informative about ability by lemma 1.

[18] This specific p_H is chosen for notational simplicity. It is the conditional probability of $a_2 = 1$ when receiving a good signal.

At the same time, $d_H(\alpha)$ is decreasing and $d_L(\alpha)$ is increasing in α, so that expected ego utility is decreasing in α. Similarly, let $c_H(\alpha)$ and $c_L(\alpha)$ be the corresponding expressions when $s_1 = 0$.

Clearly $w_0 \neq w_1$, otherwise the agent can't possibly have an incentive to observe the signal–she just chooses the option which leads to a higher (certain) payoff. Then the employer's problem is

$$\begin{aligned}
& & \min_{\alpha} \frac{1}{2}w_{a_1} + \frac{1}{2}(p_H w_1 + (1-p_H)w_0) & \\
\text{s.t. } & \frac{1}{2}u(t) + \frac{1}{2}v(w_{a_1}) + \frac{1}{2}(p_H v(w_1 + (1-p_H)v(w_0)) + \frac{1}{2}(p_H u(d_H(\alpha)) + (1-p_H)u(d_L(\alpha))) - \epsilon & \geq & \bar{u} \\
& \frac{1}{2}u(t) + \frac{1}{2}v(w_{a_1}) + \frac{1}{2}(p_H v(w_1) + (1-p_H)v(w_0)) + \frac{1}{2}(p_H u(d_H(\alpha)) + (1-p_H)u(d_L(\alpha))) - \epsilon & \geq & \\
& \max\left[v(w_{a_1}), \alpha\left(\frac{1}{2}v(w_1) + \frac{1}{2}v(w_0)\right) + (1-\alpha)(p_H v(w_1) + (1-p_H)v(w_0))\right] + u(t) & & \\
& p_H v(w_1) + (1-p_H)v(w_0)) + p_H u(d_H(\alpha)) + (1-p_H)u(d_L(\alpha)) \geq v(w_{a_1}) + u(t) & & \\
& p_L v(w_1) + (1-p_L)v(w_0)) + p_L u(c_H(\alpha)) + (1-p_L)u(c_L(\alpha)) \leq v(w_{a_1}) + u(t) & & (2.14)
\end{aligned}$$

The principal wants to minimize the expected wages to be paid out subject to four constraints. The first of these is a standard individual rationality or participation constraint. The other three are incentive compatibility constraints, making sure that the agent observes the private signal about project payoffs. IC1 means that *before* observing the signal, the agent wants to see it and condition on it rather than relying on her previous information. At the same time, IC2 and IC3 amount to saying that *after* observing signals $s_1 = 1$ and $s_1 = 0$, the agent wants to choose options 2 and 1, respectively.

We solve the problem in several steps.

1. IC1 $\Rightarrow$ IC2. (easy)

2. We ignore IC3 and will see later that it is implied by the other constraints.

3. From IC1, $\alpha > 0$ and $w_1 > w_0$.

   Intuitively, the posterior probability of the high outcome is higher after a good signal, so, in order to encourage the agent to choose option 2 in that case, the

principal has to reward $a_2 = 1$ more, just as we would expect. This is easy to formalize.

4. IR binds-otherwise the employer can just decrease all rewards, still satisfying IC1.

5. IC1 binds-otherwise the employer can offer more insurance between w_1 and w_0, slackening the IR constraints.

6. $v(w_{a_1}) = \alpha\left(\frac{1}{2}v(w_1) + \frac{1}{2}v(w_0)\right) + (1-\alpha)(p_H v(w_1) + (1-p_H)v(w_0))$

   Proof. We simply prove that neither $v(w_{a_1}) > \alpha\left(\frac{1}{2}v(w_1) + \frac{1}{2}v(w_0)\right) + (1-\alpha)(p_H v(w_1)+(1-p_H)v(w_0))$ nor $v(w_{a_1}) < \alpha\left(\frac{1}{2}v(w_1) + \frac{1}{2}v(w_0)\right)+(1-\alpha)(p_H v(w_1)+(1-p_H)v(w_0))$ is possible in an optimal solution.

   If the first one was the case, then one could offer more insurance between the outcomes w_1 and w_0 in a revenue neutral way, increasing the left-hand, but not the right-hand side of IC1.

   If the second was the case, then, since $w_1 > w_0$, we must have $w_1 > w_{a_1}$. Then decreasing w_1 and increasing w_{a_1} in a revenue-neutral way slackens both IC1 and IR. □

7. w_{a_1} is constant in α.

   Proof. Subtract IR from IC1, which both hold with equality. □

   This allows us to ignore w_{a_1} in the principal's minimization problem.

8. Without loss of generality assume that $\overline{u} = 0$. Then

$$\begin{aligned} p_H v(w_1) + (1-p_H)v(w_0) &= \Delta(\alpha) \\ \alpha\left(p_H - \frac{1}{2}\right)(v(w_1) - v(w_0)) &= \Delta(\alpha), \end{aligned} \tag{2.15}$$

   where $\Delta(\alpha) = 2\epsilon + u(t) - p_H u(d_H(\alpha)) - (1-p_H)u(d_L(\alpha))$. Solving for $v(w_0)$

and $v(w_1)$ we get

$$\begin{aligned} v(w_0) &= \Delta(\alpha) - \frac{p_H}{\alpha\left(p_H - \frac{1}{2}\right)}\Delta(\alpha) \\ v(w_1) &= \Delta(\alpha) + \frac{1-p_H}{\alpha\left(p_H - \frac{1}{2}\right)}\Delta(\alpha) \end{aligned} \tag{2.16}$$

9. The principal is interested in minimizing $p_H w_1(\alpha)+(1-p_H)w_0(\alpha)$, where we now take the wages to be functions of α. We can differentiate the above expressions for $w_0(\alpha)$ and $w_1(\alpha)$ and get

$$\begin{aligned} & p_H w_1'(\alpha) + (1-p_H)w_0'(\alpha) \\ = & \frac{p_H\Delta'(\alpha) - \frac{(1-p_H)p_H}{\alpha^2\left(p_H-\frac{1}{2}\right)}\Delta(\alpha) + \frac{(1-p_H)p_H}{\alpha\left(p_H-\frac{1}{2}\right)}\Delta'(\alpha)}{v'(w_1(\alpha))} \\ + & \frac{(1-p_H)\Delta'(\alpha) + \frac{(1-p_H)p_H}{\alpha^2\left(p_H-\frac{1}{2}\right)}\Delta(\alpha) - \frac{(1-p_H)p_H}{\alpha\left(p_H-\frac{1}{2}\right)}\Delta'(\alpha)}{v'(w_0(\alpha))} \end{aligned} \tag{2.17}$$

Let us start by examining the above derivative for two extreme cases. First, assume that the agent is (financially) risk-neutral, i.e. that v is linear. Then without loss of generality $v'(w_0(\alpha)) = v'(w_1(\alpha)) = 1$, so the derivative reduces to $\Delta'(\alpha)$, which is positive for any positive α [19]. This means that the employer wants to make α as small as possible, conditioning compensation a lot on a noisy signal of performance [20]. Although this is a highly unconventional result, in the context of this model it makes sense–since the agent doesn't care about financial risk but is averse to any real information on performance, the employer wants to drown the signal the incentives are based on in a lot of noise. With a risk-neutral agent, the principal can drown the signal in an arbitrarily large amount of noise, and by conditioning pay greatly on that noisy signal, still provide the incentives necessary.

The extensive use of stock options can be thought of as a simple practical example of the above kind of incentive structure. The performance of stocks is a very noisy

[19] It is easy to see that $\Delta(0) = 2\epsilon > 0$, $\Delta'(0) = 0$, and $\Delta''(\alpha) > 0$.

[20] For our purposes, it is not really important that for a risk-neutral agent the principal's maximization problem has no solution.

measure of the company's future profitability, let alone the performance of individual workers or the management. Therefore, if the CEO, for example, sees the stock fail, he can still reasonably convince himself that it was bad luck, and not the result of his inferior guidance. The flip side of this, and thus the drawback of using options as incentives, is that the CEO knows he has only limited influence on the stock's value, so for the incentive to work, his pay has to be conditioned on it a lot. Thus the recent examples of CEOs compensated in the tens of millions of dollars as a 'reward' for the stock market's incredible performance.

At the other extreme, when the agent doesn't care about her ego or is 'information-neutral' ($\Delta'(\alpha) = 0$), and is also strictly risk-averse, the derivative 2.17 reduces to

$$\frac{(1-p_H)p_H}{\alpha^2\left(p_H-\frac{1}{2}\right)}\Delta(\alpha)\left[\frac{1}{v'(w_0(\alpha))}-\frac{1}{v'(w_1(\alpha))}\right]<0 \tag{2.18}$$

since $w_0(\alpha) < w_1(\alpha)$. Consequently, in the optimal program $\alpha = 1$–when the agent doesn't care about her ego, we are back to the usual principal-agent problem, where adding noise to the compensation is suboptimal. This would also be the case when the agent knows her type accurately. Indeed, piece rates are common in industries where the task is so simple it is unlikely workers would attach great personal importance to doing them well, and even if they do, they can't kid themselves for very long.

For an interior optimum, and assuming a well-behaved problem, the optimal α (α^*) is the solution to the equation

$$\begin{aligned} 0 &= \left[\frac{p_H}{v'(w_1(\alpha))}+\frac{1-p_H}{v'(w_0(\alpha))}\right]\Delta'(\alpha) \\ &+ \left(\frac{1}{v'(w_1(\alpha))}-\frac{1}{v'(w_0(\alpha))}\right)\left[\frac{(1-p_H)p_H}{\alpha\left(p_H-\frac{1}{2}\right)}\Delta'(\alpha)-\frac{(1-p_H)p_H}{\alpha^2\left(p_H-\frac{1}{2}\right)}\Delta(\alpha)\right] \end{aligned} \tag{2.19}$$

This equation summarizes the basic tradeoffs of the principal. If she increases α, the principal has to pay more in expected monetary utility to the agent because the agent's expected ego utility is lower. In other words, the principal has to compensate the agent for the extra information the payoff structure forces on her. This is the first

term and tends to decrease α^*. The second two terms are related to the costliness of giving a risk-averse agent more incentives–to condition utility more strongly on the outcome while keeping expected utility the same, the principal needs to increase expected wages. The second term is the result of the fact that as α increases, it is more 'painful' for the agent to look at the signal, so the principal has to give her more incentives to do it. This effect tends to decrease α^*. On the other hand, a higher α in itself provides better incentives, since pay is more a function of actual performance. This is represented in the third term, and tends to increase α^*.

Therefore, the optimal α balances the risk- and information-aversion of the agent. Consistent with this view, it is natural to conjecture that (holding u constant) if v is sufficiently risk-averse, then α^* is close to 1, and if u is sufficiently information-averse, α^* is close to zero [21]. The following theorem implies both that the first of these statements is false and the second one is true.

Theorem 12

$$\alpha^* \leq argmin_\alpha \frac{\Delta(\alpha)}{\alpha}. \tag{2.20}$$

Proof. The system of equations 2.15 that determines $w_0(\alpha)$ and $w_1(\alpha)$ is of the form

$$\begin{aligned} p_H v(w_1(a,b)) + (1-p_H)v(w_0(a,b)) &= a \\ v(w_1(a,b)) - v(w_0(a,b)) &= b. \end{aligned} \tag{2.21}$$

It is easy to see that $p_H w_1(a,b) + (1-p_H)w_0(a,b)$ is strictly increasing in a and, since v is concave, increasing in b. In the actual system 2.15, $a = \Delta(\alpha)$ and $b = \frac{\Delta(\alpha)}{\alpha}$. Since $\Delta(\alpha)$ is strictly increasing, for any $\alpha' > \text{argmin}_\alpha \frac{\Delta(\alpha)}{\alpha}$ the principal's expected payment is greater than for $\text{argmin}_\alpha \frac{\Delta(\alpha)}{\alpha}$. □

This theorem implies that a very information-averse agent will get very noisy incentives, irrespective of her risk-aversion. So, in some sense, the information aversion

[21] It makes more sense to present this conjecture in a limit rather than a monotone comparative static way. (That is, to say instead that α^* increases as v becomes more risk-averse, etc.) Comparative statics statements are will not be true in any generality because they depend on the third derivative of v.

of the agent is a more important determinant of her incentive structure than her risk aversion is. What drives this result? Risk aversion makes using noisy signals very expensive, which should make reducing noise more important relative to protecting the agent's ego utility. However, in this problem compensating the agent for her loss in ego utility is not sufficient; if this is what the employer did, the decisionmaker would not observe her signal and just choose option 2. In order for her to choose option 2 if and only if her signal is good, she has to be rewarded more for the outcome $a_2 = 1$ relative to $a_2 = 0$, and she has to be rewarded more if the incentive structure is less noisy. This is also expensive to do for a risk-averse agent, and for a sufficiently information-averse agent the latter effect outweighs the former.

The limit this theorem sets on the informativeness of incentives depends both on the information aversion of the agent and the disutility of the task. It is easy to prove that as ϵ approaches zero, $\text{argmin}_\alpha \frac{\Delta(\alpha)}{\alpha} \to 0$, so for the easiest or most enjoyable tasks, the incentives are very noisy, no matter how risk-averse the agent. Conversely, for a large ϵ, α will be close to one.

2.6 Two Interpretations of Ego Utility

There are at least two interpretations of the reduced form utility function on beliefs that I have assumed so far in the paper. One might be called *pure self-image* and one is closer to *anxiety* or *worry* about the future. Most aspects of self-image seem to have elements of both, but the distinction is important to make.

Many of people's abilities and traits clearly affect their future earning potential, health, interpersonal relationships, and generally their enjoyment of life, in short, their utility. Thus, beliefs about our abilities mediate how we think about the future. If we derive current (hedonic) utility from thinking about our lives, these beliefs influence our current utility as well. For example, knowing that they are likely to have cancer worries most people, even though the onset of the illness, if at all, is likely to be decades away. Worrying is an unpleasant experience, so they might (depending on the functional form attached to anticipation) rather not know whether they are in

a high-risk group for the disease. In a reduced form, this is just a preference over beliefs; in addition, in some sense, such a utility function is not even a departure from traditional economics. In the model of Caplin and Leahy (1999) (based on Kreps and Porteus (1978)), people have preferences over the resolution of illusions, and having preferences over the resolution of illusions today versus later is equivalent to having preferences over beliefs [22].

This worry or anxiety aspect of self-image should be contrasted with pure self-image, which refers to the notion that we tend to care about our traits and abilities irrespective of how they will influence the future. When someone unexpectedly compliments me on my basketball game, it is unlikely that I'm happy because I anticipate how well I'll do in future games–I would probably feel equally good if I knew I would have very few chances to play in the future. It would also be a stretch to claim that I expect my life to be better because basketball ability is correlated with others that are of importance. Instead, for some reason, I care about being a good basketball player, pure and simple.

Although the emotional response to genetic susceptibility to cancer is close to the anxiety end of self-image, while many abilities and traits that constitute an integral part of our personality (like the basketball example above) are closer to the purer end, even these extremes don't fit entirely in their intended categories. Victimizing events like breast cancer, even if they are not the consequence of the patient's wrongdoing, reduce self-esteem (Taylor 1983), (Pearlin and Schooler 1978); and it can never be ruled out that what I call pure self-image is just a form of anticipation-driven self-image–the traits people care about usually influence their future utility in some way, and, worse yet, they can easily be correlated with other traits that do so even more. Hopefully, these thought exercises still demonstrate the importance of both interpretations.

Although the first part of this paper, as well as the previous chapter, focused

[22]Of course, if the agent's ability, for example, enters Caplin and Leahy's payoff space directly, then even my pure self-image model is a subcase of theirs. However, a starting point of this paper is exactly this: beliefs about an ability parameter are not generally considered to be part of the payoff space, whereas maybe they should be.

exclusively on the pure self-image interpretation of ego utility, some of its methods and results can be fruitfully exploited to study behaviors induced by anxiety as well. The following section shows this in an example motivated by the common phenomenon of people failing to go to the doctor for checkups. Due to the simpler information structure, the model below is simpler than the previous ones, but many interesting conclusions emerge nevertheless.

2.7 Anxiety

We have all at many points in our lives experienced a desire to remain ignorant: the reluctance to talk about our solutions after a test, the preferred delay in reading an important letter, or the avoidance of going to a doctor with a slight pain are all ways of suppressing unalterable information about ourselves that might, to a greater or lesser degree, influence our future. This kind of behavior is the focus of the current section.

In the area of medical psychology, there is considerable evidence that people fail to seek professional advice in time. In a study of 625 newly diagnosed cancer patients, Mor et al (1990) found that although the majority (79.5%) reported noticing symptoms prior to diagnosis, one quarter of these patients delayed longer than three months in seeking medical care. A meta-analysis of the literature on diagnostic delay indicates that 34% of women with breast cancer symptoms delay help seeking for three or more months (Facione 1993), while at the same time women are in general more likely to seek help than men. This behavior seems to be common to all demographic and social groups (Mor et al (1990)) and different personality types (Wool 1986). In some cases, it even seems that higher risk groups are less likely to take advantage of professional help: women over 50 are less likely to follow screening guidelines for cervical cancer (Simoes et al (1999)), although the incidence of the disease in this group is much higher than in younger women (Division of Cancer Prevention and Control 1986) [23].

[23]Moyer and Levine (1998) criticize the way in which many denial studies treat the concept of

It should come as no surprise that my methodology can explain this kind of behavior in a simple manner: if patients anticipate being worried about their recovery if they found out negative news about themselves, they might delay seeking help despite the fact that this help ultimately improves their prospects. A simple way to formalize this is to assume that the agent's anxiety depends negatively on the probability that she will be healthy next period: $u(Prob(d = 0))$, where $d = 0$ denotes the event of no disease [24], and u is increasing. By going to the doctor, the agent finds out more about this probability, but since the disease can be treated, it will be (weakly) higher in every state of the world. If u is concave, the agent trades off feeling better today through ignorance with getting better tomorrow through treatment.

This formulation has an immediate implication, a consequence of the expected utility formulation of anxiety. Since u is increasing, it is differentiable almost everywhere, which means that for almost every initial belief, the agent's expected anxiety loss is second order for small amounts of information. If the gains from going to the doctor and getting diagnosed are first-order, people will go to the doctor with small things. Thus, anxiety will keep patients away from seeking help only in serious cases, exactly the opposite of what standard economic logic would predict!

Section 2.4 allows us to a say a little more about the types of doctor's visits people will tend to avoid. We know that both information-averse and information-loving agents dislike news that will with some small probability reveal something very bad about them. Thus, people will not get checkups for devastating diseases which they are not likely to be diagnosed for at any given time, genetic disorders which predispose them for some illnesses, and physicals that can necessitate painful and drastic changes in lifestyle [25]. In contrast, everyone will be willing to take a test

---

denial, noting that there is no operational definition of the concept. Although they suggest future studies should be more careful in their methodology, they don't claim the results of old studies to be false.

[24] This utility function has similarities with Caplin and Leahy (1999), where it is defined over the probability that a certain type of operation will be performed.

[25] For the last example, the pure self-image and anxiety aspects of ego utility reinforce each other in an interesting way. Those who have chosen a potentially unhealthy lifestyle will not only be afraid to go to the doctor for anxiety reasons, but also because a negative diagnosis will make their earlier decision look bad (see the sluggishness result in section 2.2.3). Thus, everyone might be reluctant

where the result is likely to be negative, but possibly very favorable. This is the case when getting a second diagnosis after a bad first one, or when testing for rare immunity to a disease.

Some of the results in section 1.3 are also relevant here. If people have an abundance of voluntary opportunities to learn about their health (a reasonable assumption), in an average situation they will be averse to finding out new things about it. This has implications for what kinds of precautions they are willing to take to protect their health. In particular, section 1.3 predicts that, other things being equal, patients should be averse to precautions that are informative relative to the alternative. That is why they don't like doctor's visits–the value of an examination is fundamentally tied to its informativeness. But the model also predicts that people might not take extremely cheap home health tests like checking breasts for lumps. On the other hand, people should have no problem with cheap uninformative precautions like washing their hands or putting on sunscreen [26].

Consider also how the logic of the fear of finality applies to our example of visiting the doctor. If I notice a serious symptom on myself, should I go to the doctor? The current section argues that I might be afraid of going because finding out something bad would make me feel terrible. But let's put a twist on this: suppose I already feel terrible, knowing that the symptom is serious. Then, the doctor can either confirm my fear, in which case I'll still feel bad, but the disease can be treated, or she will disconfirm it, which would be great. Fear of finality implies that I might still not go to the doctor, but instead hope for the symptoms to go away by themselves, even if only temporarily. In that way, I'll be able to feel better, which wouldn't be possible if the doctor finalized my suspicion.

Finally, we consider a government regulation problem tied to above health psychology issue. For this, we assume that u is twice differentiable and strictly concave. Suppose there is an underlying health parameter s descriptive of the population (for

to get a lung X-ray, but smokers should be even more so.

[26]Similar comparisons can be made with regard to worker safety. Workers might refuse to wear dosimeters at a nuclear plant, while they should be willing to put on a hard hat at a construction site. The former is clearly more informative.

example, information about the incidence rate of a disease) that individuals won't monitor themselves because of anxiety. After observing s, the government can choose to legislate a stricter health policy that has an instrumental value $I(s)$ for the representative agent in preventing disease. We assume that a higher s is good news and that an intervention is less valuable for a more healthy population. More precisely, $Prob(d = 0|s)$ is increasing in s and $I(s)$ is differentiable with $I'(s) < 0$. The instrumental utility enters additively to anxiety in the representative agent's utility function [27]. We assume that the government can *not* credibly announce s itself–otherwise the problem would unfold [28]. That is, the government's actions are restricted to either legislating the new policy or not. The government is assumed to know and respect the representative agent's ego utility (anxiety). Although there are forms of utility that should probably not be included in the social welfare function, these seem to be negative social preferences (envy, hate, etc.) that run counter to the spirit of social welfare maximization. I don't see any reason why anxiety should be excluded. So if the government also cares about agents' anxiety when choosing the policy, what will it do?

Formally, this problem is somewhat similar to that considered in section 2.4. We start with the government's problem when it can commit to a legislation policy that is a function of s. Let s^* be the signal below which the government would announce a new health policy if it didn't have to worry about the population's anxiety. Focusing on policies that depend only on whether the observed s is greater than a given s^c [29], we are interested in how the government's choice of the optimal value of s^c is modified because of the latter consideration. As a function of s^c, the expected ego utility of

[27] In this formulation, the probability on which anxiety depends is not a function of whether the health policy is implemented. This unrealistic assumption is for simplicity only: the gains from anxiety reduction through a health policy can be included in I.

[28] Supposing there is a set of values of s with measure greater than zero which the government doesn't announce, there is a subset of non-zero measure of values for which $Prob(d = 0|s)$ is greater than the probability of disease conditional on no announcement. So if the government cared about anxiety, it would choose to announce those values of s if it could credibly do so.

[29] These are optimal if the function $I(s)$ is steep enough.

the representative agent is

$$Prob(s < s^c)u(Prob(d = 0|s < s^c)) + Prob(s > s^c)u(Prob(d = 0|s > s^c)) \quad (2.22)$$

The government is maximizing the sum of the ego and instrumental utilities of the agent. Thus the first-order condition for s^c is

$$\begin{aligned} 0 &= u'(Prob(d = 0|s < s^c))(Prob(d = 0|s = s^c) - Prob(d = 0|s < s^c)) \\ &+ u'(Prob(d = 0|s > s^c))(Prob(d = 0|s > s^c) - Prob(d = 0|s = s^c)) \\ &+ u(Prob(d = 0|s < s^c)) - u(Prob(d = 0|s > s^c)) + I(s^c) \end{aligned} \quad (2.23)$$

As in section 2.4, for a sufficiently information-averse u this problem is well-behaved. This means that an interior solution will be the unique root of the first-order condition 2.23.

From the same graphic interpretation as before, we can make qualitative recommendations on how s^c should be chosen. If s^* is relatively low (that is, when legislation is not likely to be necessary), then the ego part of the derivative 2.23 is negative, so the optimal s^c is lower than s^*–the government imposes health standards less often than would be warranted by instrumental considerations alone. This is because the 'legislation' signal is bad news for the population, and the government wants to reduce the probability that that signal has to be given. In plainer terms, even if health legislation seems appropriate, the government might be reluctant to mandate it because it might create a 'panic' among the people.

If s^* is relatively high (when legislation is likely to be needed), then the derivative 2.23 is positive, so s should be chosen to be greater than s^*–the government should overlegislate health measures. By doing so, the agents know that having safety measures doesn't mean much (they are legislated even when they are not really necessary), so they won't feel so bad when they see them.

The above discussion assumes an interior solution–that the government wouldn't want to follow a degenerate strategy of either mandating health measures for sure or

not doing so. This might be a reasonable strategy in certain cases. When legislation will almost certainly be required, for example, one might as well commit to getting it for sure. In these cases, though, the government's role might be superfluous: the agents could just take measures themselves without having to look at any information.

However, it is unlikely that the government could commit to a particular policy, knowing that the population won't want to check s themselves. Suppose the government has announced and people think that safety legislation will be brought in for $s < s^c$, and that we have $s \approx s^*$. In terms of instrumental utility, the government is approximately indifferent between mandating new health measures and not. However, she knows that ego utility will be strictly lower if she mandates. Therefore, she won't do so. In particular, if the optimum involved *overlegislation* of measures as described above, it's clear they won't follow their announced policy. Expecting this, the population won't believe it.

What I'm outlining is a psychological game a la Geanakoplos, Pearce, and Stacchetti (1989), where the so-called psychological equilibrium is given by correct beliefs about others' strategies and optimal behavior given others' beliefs. No matter how the population interprets the two kinds of policies, the government wants to legislate new measures if the instrumental utility of doing so outweighs the ego disutility. Therefore, the government follows a cutoff policy for any expectations of the representative agent. Under rational expectations, the cutoff value s^{nc} below which safety measures are legislated satisfies

$$u(Prob(d = 0|s > s^{nc})) - u(Prob(d = 0|s < s^{nc})) = I(s^{nc}), \tag{2.24}$$

or

$$0 = u(Prob(d = 0|s < s^{nc})) - u(Prob(d = 0|s > s^{nc})) + I(s^{nc}). \tag{2.25}$$

First, note that the left-hand side of the rational expectations outcome condition 2.25 is always positive, so, as opposed to the commitment case, safety measures are *always* legislated less often than instrumental considerations would warrant ($s^{nc} < s^*$.)

But something more interesting is true as well. Since the right-hand side of con-

dition 2.25 is less than the right-hand side of 2.23, the well-behavedness property of the commitment problem implies that if the no-commitment problem has an interior solution, then $s^{nc} < s^c$.

Theorem 13 *For any psychological equilibrium outcome s^{nc}, we have $s^{nc} < s^c$.*

This result says that even compared to what would be optimal to do taking into account the workers' ego aspect of utility, the government mandates health measures too rarely. It does so because once it is time to announce the actual policy, a benevolent social planner will take into account only the effect of her message on ego utility and the instrumental value of the policy, and not how her strategy affects the interpretation workers attach to different policies. Ex ante, this is suboptimal. In a provocative way of putting it, the social planner fails to maximize the population's utility exactly because she is benevolent–if she wasn't, she wouldn't have a reason to change her policy ex post [30].

In addition, while the commitment problem is well-behaved under sufficiently strong information aversion, there is no reason for the no-commitment problem to be the same [31]. Therefore, we might have multiple equilibria or no equilibria at all. In other words, the problem can unfold so that the government in equilibrium won't condition its legislation on new information. Once again if this is the case, its role might become superfluous as the individuals in the economy can act based on their original information as well.

The consequences of theorem 13 potentially go way beyond just health legislation. It can arise in any problem where the signal involved in the government's action affects people's utility directly. For example, consider a potential (but low-probability) disaster situation that can be prevented easily by taking some action, or an upcoming economic trouble that requires a strict economic policy. Both preventative measures

[30]The government might want to try to delegate the decision to an institution that doesn't (directly) care about the population. This has its drawbacks as well: for example, collusion between insurance companies and the institution in question.

[31]$u(Prob(q = q_H|s > s^{nc})) - u(Prob(q = q_H|s < s^{nc}))$ can easily be decreasing, *especially* for a 'very concave' u.

signal to people that there is trouble on the way. Theorem 13 says that this results in the government taking the preventative measure too rarely, a scenario endlessly played on in popular movies. Knowing this, if the government does take the action after all, the population could interpret the situation to be so bad that the behavioral consequences could be disastrous in themselves.

2.8 Conclusion

Building on the foundations of the psychology literature and arguments in chapter 1 in favor of a model based on ego utility, this paper considers further implications of a concern for self-image for behavior. If agents acquire information about their ability indirectly through making judgments and observing their quality, and they have a self-image protection motive, they will be reluctant to observe combinations of signals that involve subjective judgments. Depending on whether a later or earlier signal is more valuable, this can lead them to sluggishness or procrastination, that is, not responding to new information or delaying making a subjective judgment, respectively. It can also lead to a refusal to make decisions based on subjective judgments in the first place, relying instead on inferior objective information. In presenting applications to project choice by managers and intrinsic motivation, I also discuss how employers might try to alleviate the problems caused by self-image protection.

It seems that one of the most extensive potential applications for this kind of framework is to anxiety, which is formally the same as utility from self-image. As mentioned above, anxiety has implications for the health-related behavior of individuals, but it also complicates the analysis of optimal communication between an informed individual like the government or a doctor and an uninformed party with anxiety; in general, the mere existence of superior information can break down optimal communication, even if the interests of all parties involved are the same. How to design optimal communication channels between doctors and patients, for example, is an exciting and complicated direction for research. Another interesting area is multi-directional communication, since the doctor doesn't always have superior infor-

mation, but often needs to elicit a concern from the patient; in addition, there may be other concerned parties (relatives or friends) present. I intend to explore these questions in future research.

Chapter 3

Quasi-Hyperbolic Discounting and Retirement (with Peter Diamond)

3.1 Introduction

If you are one of the vast majority of people who think they are saving *too* little of their income for retirement[1], the natural conclusion is that you have self-control problems. If, in addition, you argued to yourself that saving more today would only lead to spending more tomorrow, and thus there is no point in saving for retirement, at least there is a small consolation: you are a *sophisticated* decisionmaker with self-control problems. And self-control problems can extend beyond savings decisions. A thirty-something Italian one of us met in Prague, had decided that it wasn't worth looking for a job anymore, because even if he got himself to do it and found one, he would quit shortly thereafter, anyway.

It is exactly these kinds of agents our paper is mostly concerned with: people who have self-control problems but realize this and behave according to it. A very clean way to model such actors is through the introduction of quasi-hyperbolic discounting[2]. This form of discounting sets up a conflict between the preferences of different intertemporal selves, and thus introduces a need for self-control. With assumptions of no commitment and that the agent takes into account her self-control problem, savings decisions can then be modeled as an equilibrium in a sequential game played by the different selves. This modeling paradigm avoids the common connection made between preference changes and cognitive failures[3], and is therefore closer to standard economic analysis. The agent in the model understands perfectly the consequences of her actions, and acts optimally within the constraints imposed by her discount function, which the psychological evidence seems to support at least some of the time[4], and the absence of easily available commitment[5].

[1]Bernheim (Bernheim 1994) reports that people 'admit to' saving much less for retirement than they should. We don't know, though, how prevalent this is among academics.

[2]Quasi-hyperbolic instead of psychologically more accurate hyperbolic discounting is used only for computational tractability.

[3]For example, Mischel and Staub (Mischel and Staub 1965) find that subjects fail to understand the contingencies involved in a decision about delay of gratification. See Ainslie and Haslam (Ainslie and Haslam 1992) for further references.

[4]E.g., Ainslie (Ainslie 1975).

[5]Admittedly, we are putting a somewhat unconventional twist on the discounted utility model. The DU model can only be justified if we put the agent's decision process entirely on the cognitive

Laibson (Laibson 1996) analyzed actors of the above kind in detail. His key result is that sophisticated actors with a quasi-hyperbolic discount structure undersave; that is, all intertemporal selves could be made better off if all of them saved a little bit more. The reason for undersaving is that from the point of view of self t, self $t+1$ consumes too much, in essence wasting part of the savings inherited from self t. Self t doesn't like that, and she can do nothing about it, so she just gives self $t+1$ less. Both would love to commit self $t+1$ to saving more (self t because her savings wouldn't be wasted, and self $t+1$ because she would receive more savings), but the technology is simply not available.

We adapt Laibson's basic setup for the analysis of the effect of endogenous retirement decisions on savings behavior. The addition will be simply that in each of the models there will be a single period in which the agent can choose whether to work or retire. Working will cost the agent some utility, but she will be compensated with extra wealth. Commitment will not be possible: agents cannot precommit to a decision concerning retirement, nor to any consumption level.

We are mostly interested in building intuition through examining how the interaction with the retirement decision changes the Laibson savings problem. There are various benchmarks that involve the Laibson-type dynamics and that arise naturally in at least one of our models. All of them involve taking away the choice concerning retirement in one form or another. First, we can contrast equilibrium savings levels to those that would arise if there was a prior commitment possibility regarding work (delayed retirement), but no possibility of commitment about savings. At other times, a more useful comparison will be the simple Laibson-type equilibrium that results when the transition to retirement is determined exogenously, as, for example,

---

plane: unless we define utility in Baron's sense as the fulfillment of goals (Baron 1993), which makes the definition of an instantaneous utility function extremely problematic, we can't say the decisionmaker experiences discounted utility in any way. The consequence of hyperbolic or quasi-hyperbolic discounting, namely, a conflict in the cognitive plane–where information is processed and decisions are made–then seems to imply irrationality on the part of the agent. This, however, is only so if we believe the understanding of humans as possessing an unchanging, non-contradictory essence. If the essence is changing or contradictory (that being outside our realm of choice and therefore not a question of rationality), that will translate into conflicts for our decision-making part. And in as much as self-control problems are exactly about the non-existence of such an entity, we definitely don't want to assume it for a study of them.

with a mandate. The natural question to start with is whether a lack of commitment (or choice) results in higher or lower savings levels for retirement.

We will start with the simplest model that is relevant in (quasi-)hyperbolic discounting: the three-period model in which the middle period is the retirement decision period. Self 2 values the payoff compared to the effort she has to exert much less than self 1, so she will sometimes retire when self 1 would like her to work. In order to avoid this outcome, self 1 might save less (than she would if she could commit self 2 to work) to make self 2 work. On the other hand, if self 1 would like self 2 to work, but it is too expensive to achieve that without commitment, she will save more to help finance self 2's unavoidable retirement (than if she could commit self 2 to work). Note the qualitative distinction between a change in self 1's saving (compared to a setting with commitment) to induce a retirement decision and to accommodate one. Here, we can get lower saving to block the 'threat' of retirement and higher saving to accommodate it.

These effects are unchanged qualitatively by the addition of labor effort uncertainty or a longer horizon after retirement. Things get much more complicated when we allow for more periods before retirement. If lower saving is the outcome, it will be split by the pre-retirement selves–since the Laibson equilibrium is already characterized by overconsumption, earlier selves don't like to leave the job of undersaving to later selves. However, at least two periods before retirement new effects can also come into play. In particular, there can be a conflict between earlier and later selves about too *late*, not too early, retirement. This is because with quasi-hyperbolic discounting successive selves agree in what the later selves should do, they just don't agree on how much it is worth to induce them to do it. And the earlier self will always prefer higher savings levels than the later self. Thus we can get higher saving to 'encourage' early retirement. This effect exhibits much more mixed global implications than the lower saving one, mostly because earlier selves don't necessarily want to take part in the higher saving.

In the presence of a retirement decision there are a number of ways to observationally distinguish quasi-hyperbolic and exponential discounting. As Laibson has noted,

in the savings game the path of consumption can't be used to distinguish the two, only some comparative statics observations can be (Laibson 1996). This is not true if there is a retirement choice parameter: if higher saving or lower saving (which are not possible with exponential discounting) happens, the consumption path won't look like that of an exponential discounter. In addition, there are numerous changes in the economy to which an agent endowed with quasi-hyperbolic discounting will react differently. The most radical diversion from the predictions of consistent preference models emerges when we consider the effect of an increase in wage level in the endogenous retirement period. If the agent undersaves to make the deciding self work, and the need for lower savings to induce work is relaxed through higher earnings, she will save more, giving a negative marginal propensity to consume out of changes in future income.

We realize, however, that many, if not most, of us fail to grasp fully the extent of our self-control problem. That is, we often fail to be fully sophisticated in the sense of the Laibson model. Therefore, we will briefly discuss the potential outcomes under the particular assumption of naiveté, that each self falsely assumes that the others will comply with her plans. Since there is no game in this case, the analysis is considerably simpler. One interesting implication of naiveté is the possibility that the selves before the deciding self plan to retire late, but the deciding self chooses to retire early, leading to an update in lifetime wealth and thus a drop in the consumption path at retirement.

3.2 The quasi-hyperbolic discounting setup

We adapt the structure recently used by Laibson for analyzing quasi-hyperbolic discounting issues. For a more detailed introduction, see (Laibson 1996), for example. The consumer's instantaneous utility function is of the constant relative risk aversion (CRRA) class, that is

$$u(c) = \frac{c^{1-\rho}}{1-\rho} \text{ if } \rho \neq 1, \text{ and } u(c) = \ln(c) \text{ if } \rho = 1, \tag{3.1}$$

ρ being the risk aversion parameter. A nice property of CRRA utility functions is the fact that for intertemporal maximizations of the form

$$\begin{aligned} &\max_{c_1,c_2} u(c_1) + \kappa u(c_2) \\ &\text{s.t. } c_1 + \frac{1}{R}c_2 = W \end{aligned} \tag{3.2}$$

with κ a positive discount factor, the solution will always be $c_1 = \lambda(R,\kappa)W$ for some $0 < \lambda(R,\kappa) < 1$. Also, then, some easy manipulation shows that lifetime discounted utility can be written as $K(R,\kappa)u(W)$ (or $K(R,\kappa) + u(W)$ for $u(c) = \ln(c)$) for a positive function $K(R,\kappa)$. This allows us to collapse periods where we have already solved the problem and gotten linear answers into a single period, a shortcut extremely convenient for backward induction arguments. We will use this property a number of times in the paper.

In a T-horizon game, self t's discounted utility from present and future consumption is

$$u(c_t) + \beta \sum_{i=1}^{T-t} \delta^i u(c_{t+i}) \tag{3.3}$$

with an expectation at front if there is uncertainty. β and δ (both between 0 and 1) are discount parameters meant to capture the essence of hyperbolic discounting, namely that the discount factor between adjacent periods close by is smaller than between similar periods further away. Indeed, the discount factor between periods t and $t+1$ is $\beta\delta$, and between any two adjacent periods later it is δ.

Of course, the discount structure just described applies only to self t; for example, self $t+1$'s discount factor between $t+1$ and $t+2$ is $\beta\delta$. Therefore, there is a conflict between different selves about how much to consume (or whether to retire) in a given period, or, more formally, preferences are intertemporally inconsistent. We assume that commitment is not possible (so that each self controls her period's consumption, subject to a financial or wealth constraint[6], and possibly a decision concerning retirement), and model the behavioral decisions as a subgame-perfect

[6]depending on whether there are liquidity constraints.

equilibrium of the game played by the different selves[7]. Finally, R is the constant and exogenous gross return on wealth.

3.3 The three-period model

We begin with the three-period model, the shortest possible that actually generates quasi-hyperbolic discounting effects. The periods are labeled 1,2,3, and subscripts on c or W refer to the period in question. In the first period, the agent has to work; in the second, she can decide whether to work or retire; and in the third, she has to retire[8]. The agent incurs a constant utility cost of effort $e > 0$ if she works in the second period, but she also gets an extra Δ amount of income if she does.

As usual when looking for subgame-perfect equilibria, we solve backwards. The decision is easy in the third period: no work is done and all remaining wealth is consumed. Suppose, then, that the period 2 self inherits a wealth of W_2. This will be her remaining wealth if she retires, and she will have $W_2 + \Delta$ if she works. As we have mentioned above, there is a $\lambda > 0$ such that self 2 will always consume a proportion λ of her wealth. Thus her discounted utility is

$$u(\lambda W_2) + \beta\delta u(R(1-\lambda)W_2) \tag{3.4}$$

if she doesn't work, and

$$u(\lambda(W_2 + \Delta)) + \beta\delta u(R(1-\lambda)(W_2 + \Delta)) - e \tag{3.5}$$

[7]The game theory-based decision rule is basically equivalent to the assumption of sophistication on the part of the agent. An alternative assumption is naiveté, where each self naively assumes that others will follow her decisions. We will study naifs briefly in section 3.7.

[8]Making the key period–that of the retirement decision–the first or third period takes all the spice out of the model because then the conflict to be described below wouldn't materialize.

if she works. Therefore she will work iff[9]

$$u(\lambda(W_2+\Delta)) - u(\lambda W_2) + \beta\delta u(R(1-\lambda)(W_2+\Delta)) - \beta\delta u(R(1-\lambda)W_2) \geq e. \quad (3.6)$$

Since u is concave, there is a $\overline{W}_2$ such that self 2 will retire iff $W_2 > \overline{W}_2$.

Now let's look at this situation from the point of view of self 1. She will prefer self 2 to work if

$$\beta\delta u(\lambda(W_2+\Delta)) - \beta\delta u(\lambda W_2) +$$

$$+\beta\delta^2 u(R(1-\lambda)(W_2+\Delta)) - \beta\delta^2 u(R(1-\lambda)W_2) \geq \beta\delta e,$$

or

$$u(\lambda(W_2+\Delta)) - u(\lambda W_2) + \delta u(R(1-\lambda)(W_2+\Delta)) - \delta u(R(1-\lambda)W_2) \geq e. \quad (3.7)$$

Notice that the left-hand side of 3.7 is greater than the left-hand side of 3.6; consequently, there is a range of wealth levels for which self 2 won't work, but self 1 would like her to. In particular, for $W_2 = \overline{W}_2$ self 2 is indifferent between working and not working, but self 1 strictly prefers her to work. This effect arises simply because self 1 weighs the cost and the benefit of working differently: for her, the cost is less salient.

Figure 3-1 displays the continuation utility for self 1 (her utility from periods 2 and 3) as a function of W_2, the level of wealth self 1 leaves for self 2, for an example with logarithmic utility. The curve that starts off as a solid line and continues as a dotted one (U_w) is self 1's utility *assuming* self 2 works, and the other curve (U_r) is her utility assuming self 2 doesn't work. Only the solid part of each curve is available to self 1, as she has to take into account self 2's decision. Nevertheless, the simplest way to understand self 1's maximization problem is through U_w and U_r. Define s_i^* (i being r or w) to be the wealth received by self 2 in the solution to the maximization

[9] We are assuming for now that the agent will work if she is indifferent. In the long-horizon models, we will more generally assume that an agent indifferent between two actions will choose the one the earlier selves would prefer. (With quasi-hyperbolic discounting, all earlier selves want the same thing.) It turns out that this gives the essentially unique subgame-perfect equilibrium–otherwise, the earlier self's maximization problem has no solution.

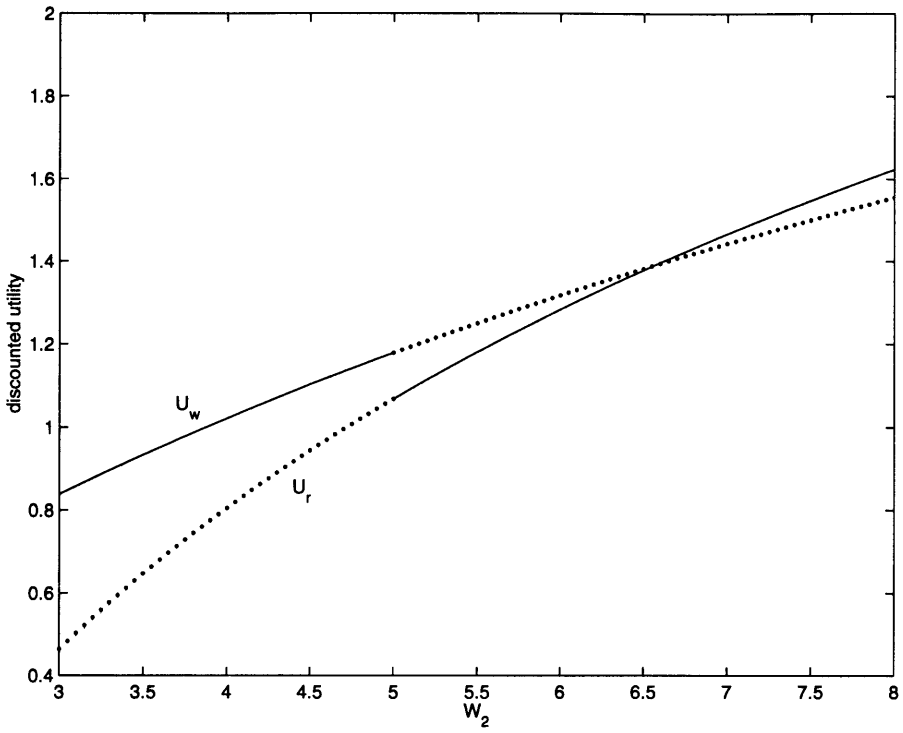

Figure 3-1: Utility of self 1 from periods 2 and 3 with and without work in period 2

problem that ignores the endogeneity of retirement:

$$\max_s u(W_1 - \frac{1}{R}s) + U_i(s). \tag{3.8}$$

Note that $s_w^* < s_r^*$ since work provides extra income in period 2, and some of that is consumed in period 1. If self 1 could commit self 2 to a decision on work (but not on consumption), she would choose one of these savings levels. Therefore, let s_c^* be the better of the two possibilities s_i^*. That is, s_c^* gives the optimal savings level if self 1 could commit self 2's retirement decision. We describe optimal savings levels by examining these constructs.

If self 1 would commit self 2 to retire, that is, $s_c^* = s_r^*$, then $s_r^* > \overline{W}_2$, the level of wealth at which self 2 is indifferent to retirement. Otherwise s_r^* would be dominated by a working alternative. Consequently, in that case the optimal savings level is $s^* = s_c^* = s_r^*$; an inability to commit to retirement does not matter if self 1 wants

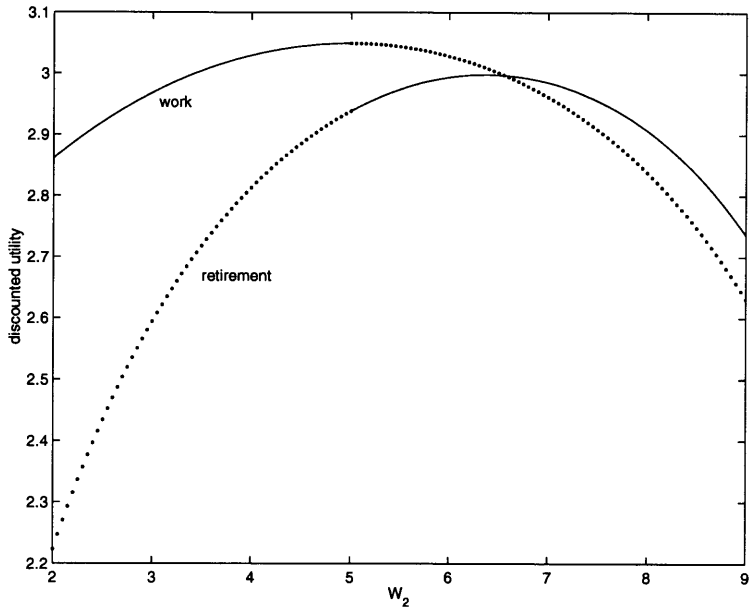

Figure 3-2: Lifetime utility of self 1

self 2 to retire. If $s_c^* = s_w^*$, things take a more interesting turn. If $s_w^* \leq \overline{W}_2$, then analogously to the above $s^* = s_c^* = s_w^*$. However, it is also possible that $s_w^* > \overline{W}_2$, so that s_c^* and work in period 2 is not available to self 1. Then–since U_w is concave–the best point on the available part of U_w is the boundary, $s_b^* = \overline{W}_2$. This might or might not dominate the best point resulting in retirement, which satisfies $s_r^* > s_w^* > \overline{W}_2$. Thus, two possibilities emerge: either $s^* = s_b^* = \overline{W}_2 < s_w^* = s_c^*$, or $s^* = s_r^* > s_w^* = s_c^*$. Notice that in each of these cases the equilibrium savings level is different from the one under commitment.

Figure 3-2 shows lifetime discounted utilities for self 1 as a function of W_2 assuming work and retirement in period 2 for the same example as in figure 3-1. Again, the solid part of each curve is available to self 1. s_w^* maximizes the work curve, s_r^* the retirement curve, and, as we have seen, the best available point among s_w^*, s_r^*, and $\overline{W}_2$ is the optimum for self 1. In this example, it seems to be $\overline{W}_2$.

We can trace out the equilibrium as a function of W_1 for the general case. For low values of W_1, $s_w^* < \overline{W}_2$, self 2 works in equilibrium and the inability of self 1 to commit self 2 to work has no effect. Then, there is a range of values for W_1 such that

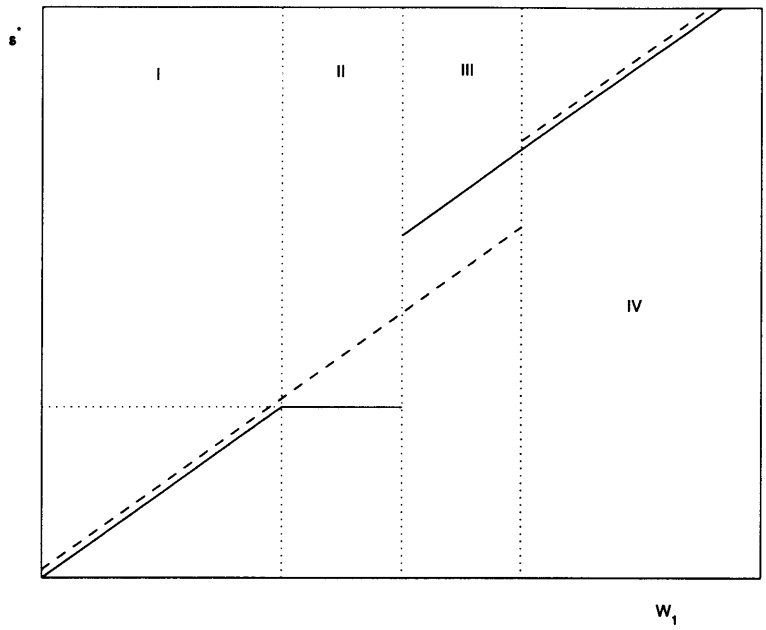

Figure 3-3: Savings of self 1 with (dashed line) and without (continuous line) commitment

Note: the figure is only qualitative; it is not meant to illustrate actual slopes or relative sizes for the regions.

optimal savings equals $\overline{W}_2$ in order to just induce work. Over this range savings are less than they would be if self 1 could commit self 2 to work. In the next range of W_1, self 1 accommodates self 2, saving $s_r^* > \overline{W}_2$ even though self 1 would save less and commit self 2 to work if that were possible. For high enough values of W_1, self 1 prefers that self 2 retire and there is, again, no effect from the inability to commit. This is shown in figure 3-3.

The marginal propensity to consume in period 1 out of a small increase in W_1 behaves differently in the different regions[10]. In the lowest region, for a small increase in W_1, the fraction of the increase consumed is λ_1, just as in the case without a retirement decision. For a small increase in wealth in the second region, all of it is consumed so that self 2 continues to receive $\overline{W}_2$. For small increases in wealth in the top two regions, again, the fraction λ_1 is consumed in period one.

[10]By a small increase we mean one that does not move self 1 into a different region.

When interpreting results in these short-horizon models, we have to be very careful not to confuse genuine quasi-hyperbolic discounting effects with effects that arise due to the fact that we have chosen a short horizon. In particular, you might notice that even if $s^* = s_i^*$ for an i, we don't have $\frac{c_1}{c_2} = \frac{c_2}{c_3}$ as we do for exponential discounting with CRRA utility functions. But this peculiarity occurs only because the marginal propensities to consume change from period to period, a property that disappears as the horizon after retirement is assumed to go to infinity[11]. In that case, only the equivalent of $s^* = \overline{W}_2$ will not satisfy the equivalent of $\frac{c_1}{c_2} = \frac{c_2}{c_3}$. But you don't have to turn to section 3.5 to understand that the case $s^* = \overline{W}_2$ is the only one observationally different from exponential discounting: it is the only case when self 1 uses non-optimal savings (in the sense of the consumption game) as a tool to change the retirement decision of self 2. And as Laibson has pointed out in the context without a retirement decision, optimal savings with quasi-hyperbolic discounting is observationally equivalent to exponential discounting (Laibson 1996). Non-optimal savings, finally, is not possible with exponential discounting, even in the presence of a retirement decision: in that case $\overline{W}_2$ is defined by the intersection of the curves U_w and U_r, so s_w^* and s_r^* both dominate it, and s_c^*, which is always available, does so strictly.

Behaviorally, as opposed to just observationally, there is another, more subtle, difference between quasi-hyperbolic and exponential discounting. It is possible that self 1 would prefer to commit self 2 to work and give her $s_c^* = s_w^*$, but since that is not possible and she has to undersave too much to make self 2 work, she chooses $s^* = s_r^*$. This reason for choosing s_r^*, though unobservable, *is* unique to quasi-hyperbolic discounting: it arises from the conflict of self 2's decisions and self 1's wishes. However, the reason for higher saving in this case (clearly $s^* > s_c^* = s_w^*$) is very different from

[11] The marginal propensity to consume matters with quasi-hyperbolic discounting simply because the Euler equation contains it:

$$\frac{u'(c_t)}{u'(c_{t+1})} = R\delta\left(\beta\frac{\partial c_{t+1}}{\partial W_{t+1}} + 1 - \frac{\partial c_{t+1}}{\partial W_{t+1}}\right).$$

This is proved in (Laibson 1996) but also falls out as a special case of our analysis in section .6 of the appendix.

the reason for lower saving above[12]: it is not intended to change the retirement decision of self 2. Quite the opposite: in recognition of the fact that it would be 'too expensive' to change self 2's decision, self 1 will save more to offset the lower wealth level of self 2 due to the early retirement.

This is a point where the addition of a retirement decision fundamentally changes the Laibson problem. With only savings, the 'wasting' of wealth on the part of later selves always decreases the marginal utility of savings for the earlier self (that's why that self saves less). When there is a retirement decision, however, a 'wrong decision' by the later self can decrease wealth, and thus increase the marginal utility of savings[13].

Though very simple, the three-period certainty model delivers much of the intuition that we will encounter in more complicated models. In the next section, we show that the different cases discussed here are robust to the addition of uncertainty. Adding uncertainty will eliminate case analysis and transform it into effects analysis, thereby also allowing a delineation of when higher or lower saving is likely to occur.

3.4 Uncertainty

We could introduce uncertainty in period 2 labor income (Δ), and in period 2 cost of effort (e). The two give similar results, and the latter is somewhat nicer to present, so we present only that one. Assume therefore that self 1 doesn't know e, but knows its continuous distribution function f (and the cumulative distribution function F). This standard assumption is made plausible by the possibility that the agent does not know how healthy or how thrilled she will be to work in the future. We assume that the support of f is wide enough to encompass all of the regions above.

We start again from self 2's problem, who has inherited a wealth W_2. Define

[12]Remember that the benchmark for all savings discussions at this point is the savings level that would arise if self 1 could control self 2's retirement decision.

[13]The $s^* = \overline{W}_2$ case has an interpretation in the Laibson spirit: since savings would be wasted by self 2 to finance retirement, self 1 saves less to not let self 2 do that.

$\overline{e}(W_2)$ as the level of effort cost at which self 2 is indifferent to work:

$$u(\lambda(W_2+\Delta))+\beta\delta u(R(1-\lambda)(W_2+\Delta))-\overline{e}(W_2)=u(\lambda W_2)+\beta\delta u(R(1-\lambda)W_2) \quad (3.9)$$

Self 2 will work if $e \leq \overline{e}(W_2)$. Therefore self 2 will work with probability $F(\overline{e}(W_2))$ and retire with probability $1-F(\overline{e}(W_2))$. For simplicity, let K be the constant such that $\beta\delta u(\lambda W)+\beta\delta^2 u(R(1-\lambda)W)=Ku(W)$. As we have mentioned, such a constant always exists for CRRA utility functions[14]. Now the maximand for self 1 is

$$u(W_1-\frac{1}{R}W_2)+K[F(\overline{e}(W_2))u(W_2+\Delta)+(1-F(\overline{e}(W_2)))u(W_2)]-\beta\delta\int_0^{\overline{e}(W_2)}ef(e)de. \quad (3.10)$$

The first-order condition is

$$\frac{1}{R}u'(W_1-\frac{1}{R}W_2)=K[F(\overline{e}(W_2))u'(W_2+\Delta)+(1-F(\overline{e}(W_2)))u'(W_2)+$$
$$f(\overline{e}(W_2))\overline{e}'(W_2)u(W_2+\Delta)-f(\overline{e}(W_2))\overline{e}'(W_2)u(W_2)]-\beta\delta\overline{e}(W_2)f(\overline{e}(W_2))\overline{e}'(W_2),$$

which is equivalent to

$$\frac{1}{R}u'(W_1-\frac{1}{R}W_2)=K[F(\overline{e}(W_2))u'(W_2+\Delta)+(1-F(\overline{e}(W_2)))u'(W_2)]+$$
$$f(\overline{e}(W_2))\overline{e}'(W_2)[Ku(W_2+\Delta)-\beta\delta\overline{e}(W_2)-Ku(W_2)]. \quad (3.11)$$

A similar first-order condition would arise if self 1 could commit self 2 to a state-contingent retirement decision[15], except that $\overline{e}(W_2)$ should be replaced by $\tilde{e}(W_2)$, where $\tilde{e}(W_2)$ is defined by

$$\beta\delta u(\lambda(W_2+\Delta))+\beta\delta^2 u(R(1-\lambda)(W_2+\Delta))-\beta\delta\tilde{e}(W_2)=\beta\delta u(\lambda W_2)+\beta\delta^2 u(R(1-\lambda)W_2). \quad (3.12)$$

[14]When the utility function is logarithmic, the correct expression is $\beta\delta u(\lambda W)+\beta\delta^2 u(R(1-\lambda)W)=K+u(W)$. The analysis is the same in this case.

[15]A commitment device conditional on the realized e is not very realistic, but as a comparison it is useful for highlighting the tradeoffs self 1 faces. If self 1 could only commit to a specific decision (one not conditional on e), she would never commit to retirement, and to work only if that is not too costly on the high-e end.

(This just defines the cutoff cost level under which self 1 would want self 2 to work.) Then, by definition, $Ku(W_2+\Delta)-\beta\delta\tilde{e}(W_2)-Ku(W_2)=0$, so the first-order condition is

$$\frac{1}{R}u'(W_1-\frac{1}{R}W_2)=K[F(\tilde{e}(W_2))u'(W_2+\Delta)+(1-F(\tilde{e}(W_2)))u'(W_2)]. \qquad (3.13)$$

Neither of these two first-order conditions is well-behaved, and we have not found simple conditions on f that would make them well-behaved. What we would like is for the right-hand sides of equations 3.11 and 3.13 to be decreasing in W_2. Then we would have unique solutions to the first-order conditions, which would be global maxima. Notice that the derivative of the right-hand side of 3.11 is of the form

$$K[F(\bar{e}(W_2))u''(W_2+\Delta)+(1-F(\bar{e}(W_2)))u''(W_2)]+f(\bar{e}(W_2))[Z]+$$
$$f'(\bar{e}(W_2))\bar{e}'^2(W_2)[Ku(W_2+\Delta)-\beta\delta\bar{e}(W_2)-Ku(W_2)], \qquad (3.14)$$

where the expression Z multiplied by $f(\bar{e}(W_2)$ is complex and not worth writing down for our purposes. The derivative of the right-hand side of 3.13 is very similar, the difference being that $\bar{e}(W_2)$ is replaced by $\tilde{e}(W_2)$ and there is no term multiplied by f'[16]. Now it is easy to see that if f and f' are 'small enough' (though it is hard to give meaning to this phrase), the problem is well-behaved. For example, a uniform distribution with a large enough support will do. This is certainly a sufficient condition, albeit not necessary.

Having said that, we assume that unique solutions to the FOCs exist, in which case they define the maximum. We are interested in the difference of the right-hand-sides

[16]The term multiplied by $f(\bar{e}(W_2)$ is also simpler.

of the first-order conditions[17]:

$$K\overbrace{[F(\tilde{e}(W_2)) - F(\overline{e}(W_2))][u'(W_2) - u'(W_2 + \Delta)]}^{\text{higher saving}} + f(\overline{e}(W_2))\overline{e}'(W_2)\underbrace{[Ku(W_2 + \Delta) - \beta\delta\overline{e}(W_2) - Ku(W_2)]}_{\text{lower saving}} \tag{3.15}$$

Notice that since $\tilde{e}(W_2) > \overline{e}(W_2)$ for any W_2, the overbraced product is positive, so it indeed encourages higher saving. On the other hand, we know that for $\overline{e}(W_2)$, self 2 is indifferent between working and not working, and also that in that case self 1 would prefer her to work. Thus the underbraced term is positive. But $\overline{e}'(W_2)$ is negative, so the given effect in fact tends to lower savings.

The intuition behind these two effects is straight-forward enough. First, since there is a chance that self 2 will retire when self 1 prefers that she work, she'll need more money than if she worked. Thus, self 1 saves more. Second, since saving less induces work in some additional states, self 1 has an incentive to save less.

It should be clear that these are just translations of the cases analyzed in the certainty model into the uncertainty setting. This setup, in addition, also allows for convenient analysis of when higher or lower saving is likely to occur. For example, if $f(\overline{e}(W_2^c))$ (where W_2^c is optimal savings with commitment) is high compared to $F(\tilde{e}(W_2^c)) - F(\overline{e}(W_2^c))$, we will get lower saving. That is, if self 1 feels that she can exert a lot of influence on self 2's decision through savings, she will save less. On the other hand, if $f(\overline{e}(W_2^c))$ is close to zero, while $F(\tilde{e}(W_2^c)) - F(\overline{e}(W_2^c))$ is fairly large, there will be higher saving. In simpler terms, if self 1 can't exert much influence on self 2, she will just accept that self 2 might retire too early, and give the now poorer self more savings[18].

[17] If the difference is positive at the optimal savings with commitment, then the optimal savings without commitment is higher. This is trivial if the problem is well-behaved in the above sense. But the assumption that the first-order condition has a unique solution, together with the observation that for low W_2 the right-hand side of equation 3.11 is greater than the left-hand-side, and vice versa if W_2 is close to RW_1, is also sufficient. Similarly, the opposite is the case if the difference is negative.

[18] Notice that making the size assumptions on f and f' does not make the comparison of the two effects an irrelevant exercise. Though f and f' are small (compared to 1), there is no restriction on

3.5 Multiple periods before the retirement decision

In the savings game, the tendency of long horizons to make marginal propensities to consume approximately equal across periods helps both in describing the quasi-hyperbolic equilibrium and in comparing it with the exponential discounting outcome. Put differently, a long but finite horizon is a convenient tool to pin down the equilibrium while keeping the smoothness properties of an infinite horizon. But it doesn't offer many surprises. This is also the case in our model if the horizon after retirement is long. To create a 'disagreement' between the self making the work/retirement decision and the previous one, which is what drives the results of our previous sections, one only needs a setting in which the cost of extra work is concentrated in a single period, while the benefits of it are spread out. This is certainly satisfied if we have a long horizon after retirement, so as long as there is only a single period before the decision period, we shouldn't expect the results to change too much. In fact, they don't; but since a detailed discussion would add little, it is relegated to the appendix.

Though introducing a long horizon after retirement is of little consequence to the qualitative results of our models, a longer horizon before retirement does set up a novel, interesting distinction: how the effects play themselves out close to versus far from retirement[19]. Unlike in the savings game, what happens at the end is no longer an empirically unattractive theoretical nuisance; the behavior is not a response to nearby deterministic death, but to the approach of the end of working life–the central focus of this paper. Thus, we will 'move backwards' in this section, and see what happens when the horizon before retirement is let to grow. Perhaps surprisingly, the effects change considerably.

---

their *relative* size, so $f(\overline{e}(W_2^c))$ and $F(\tilde{e}(W_2^c)) - F(\overline{e}(W_2^c))$ might compare in any number of ways.

[19] As we have seen in the previous section, uncertainty helps in putting effects nicely side-by-side. Also, it seems to make the problem smoother, eliminating discontinuities in selves' strategies (Laibson 1997). However, there is only an Euler equation to work with, and for a long horizon before retirement we know little about the global properties of the solution. While this might be satisfactory in a savings problem, it is inherently unfitted for analyzing the retirement decision: the Euler equation says nothing about it. So we will have to do with the certainty model for this section.

Unfortunately, there is very little we can say about the equilibrium in general with many periods of work. The bulk of the trouble stems from the fact that when later selves have decreasing marginal propensities to consume[20], the consumption schedules become extremely complicated very quickly as we move to earlier periods.

All we know is that the agent's consumption schedule is piecewise linear in wealth for each t, and, furthermore, the agent's consumption path is as if she was going through a series of shorter Laibson problems[21]. This is quite interesting in itself: the agent periodically acts as if she is liquidity constrained and/or impatient, even though she has perfect foresight and faces no constraints. But since we are unable to say much in general about the equilibrium, we will mostly restrict our attention to a model in which there are only two periods of exogenously mandated work before the period of endogenous decision, though some results will be more general than that. For notational convenience, we now label the period with a retirement decision as period 0–the first period of life is thus period -2. We assume a long horizon after retirement, although the qualitative results are the same with a shorter lifespan[22].

Before we plunge into the work, we reconsider the benchmark for our savings discussions. For the three-period model, the difference from the benchmark case of what would happen if self 1 could commit self 2 to a retirement decision answered the question of how the conflict between selves 1 and 2 was reflected in savings. For longer horizons before retirement, the conflict is not only between the current self (say self t) and the self making the retirement decision–there are many selves in-between with whom self t may also have a conflict. An important implication of this is that self t *might not want to* commit to a retirement decision. Commitment also allows other selves to behave differently, which self t might not like[23].

[20]Laibson (Laibson 1997) describes such an example in detail, though in the context of liquidity constraints. Here, since self 2 in section 3.3 has a region where her marginal propensity to consume is 1, in that region she behaves as if 'liquidity constrained.' This gives the jumps in consumption earlier on.

[21]See the appendix for a formal statement and a proof.

[22]Also, we will assume in this section that a Markov-perfect equilibrium in pure strategies exists for the game. The proof of this claim is contained in the appendix. Noticeably, that proof uses similar methods to those below, but putting it in the appendix and just assuming existence for now makes the paper much easier to follow.

[23]We could say that we are comparing things to when self t is forced to make a commitment, but

Fortunately, this issue is not too critical if the earliest self considered is self -2, so we will keep using similar vocabulary to section 3.3, while realizing that this would be inappropriate for longer horizons.

As we have mentioned, self -1's behavior is qualitatively the same as what we have seen in section 3.3. Let us now move back to self -2 and see what she thinks about the behavior of self -1. (Notationally, we include the PDV of earnings in all periods except 0 in the wealth measure W, considering only the possible earning Δ in period zero separately.) Suppose giving the next self the savings level in a Laibson problem ($W_{-1} = R(1-\lambda^*)W_{-2} - \frac{1}{R}\lambda^*\Delta$) would result in a savings decision $s^* < s_c^*$ by self -1. In the pure Laibson problem with work in period zero, self -1 is already consuming too much from the point of view of self -2 (even the savings level s_c^* is too low for her), so an even larger consumption should leave self -2 rather unhappy. If self -2 likes the working alternative, she would rather do part of the 'overconsumption' herself. And we know she can do that, since for lower wealth levels self -1 still prefers late retirement. In fact, we have the following more general, intuitive result:

Lemma 3 *Suppose that $t \leq 0$ and $W_t > W_t'$. Then it is not possible that self t with wealth W_t behaves so that self 0 eventually works, and with wealth W_t' she behaves so that self 0 eventually retires.*

The formal proof is in appendix .4. It takes advantage of the concavity of consumption utility to show that savings is monotonically increasing in wealth for each self before zero. This implies that self 0's wealth is monotonically related to previous selves' wealth levels. And we know self 0 retires iff $W_0 > \overline{W}_0$ for a given $\overline{W}_0$.

The above result rested on the assumption that self -2 liked the working alternative. What if she doesn't? We have already seen that she views the tradeoff between the extra wealth and making self 0 work differently from self -1. But even without lower saving by self -1, she certainly views the benefit and cost of working an extra period differently. If so, what kind of conflicts arise between self -2 and self -1?

---

if that is against self t's will, the interpretation of the results is ambiguous.

The answer might be surprising: self -2 will never use boundary (knife-edge) savings to get the working alternative, but it is possible she will use it to 'force' retirement. This is exactly what the following lemma proves.

Lemma 4 *Let $\overline{W}_t$ $(t < 0)$ be the level of wealth at which self t is indifferent between behaviors that eventually lead to self 0 working or retiring. At this savings level, self $t-1$ strictly prefers self t to choose to eventually make self 0 retire.*

Once again, the proof is in the appendix, but it's essence is simple: due to the different preferences, self t cares relatively more about consumption in period t than does self $t-1$, so when self t is indifferent, self $t-1$ wants her to go for the low-consumption (high-saving) option. And this is of course the early retirement option. Self -2, then, might save more than with mandated early retirement to just induce self -1 to save so as to result in early retirement.

These lemmas can be used to illustrate self -2's general qualitative savings behavior relative to wealth, which is done in figure 3-4. For very low levels of wealth, self -2 prefers late retirement, and this can be achieved under the Laibson consumption solution.

In the next two regions (II and III), selves -2 and -1 undersave to induce self 0 to work. By lemma 3, self -2 can split the undersaving with self -1, while still inducing eventual late retirement. For relatively low wealth levels where there has to be undersaving done to induce self 0 to work (region II), all the undersaving will be done by self -2. In this region, self -1's marginal propensity to consume is 1, so self -2 prefers to consume all extra wealth as long as $u'(c_{-2}) > \beta\delta u'(c_{-1})$, and her savings function is flat as a function of wealth. As self -2 gets richer, she will want to split the extra consumption with self -1 even though self -1 has a marginal propensity to consume of 1. In this region (III), we have $u'(c_{-2}) = \beta\delta u'(c_{-1})$, and the savings function is positively sloped, although with a lower slope than in region I [24].

[24]There are some things that can be said in greater generality. Assume that each self $t < 0$ already prefers early retirement for a low enough wealth level so that there are no jumps in self t's consumption function on the late retirement section. (We expect the statements that follow to be true even without this assumption, but haven't been able to prove it.) Then it is easy to

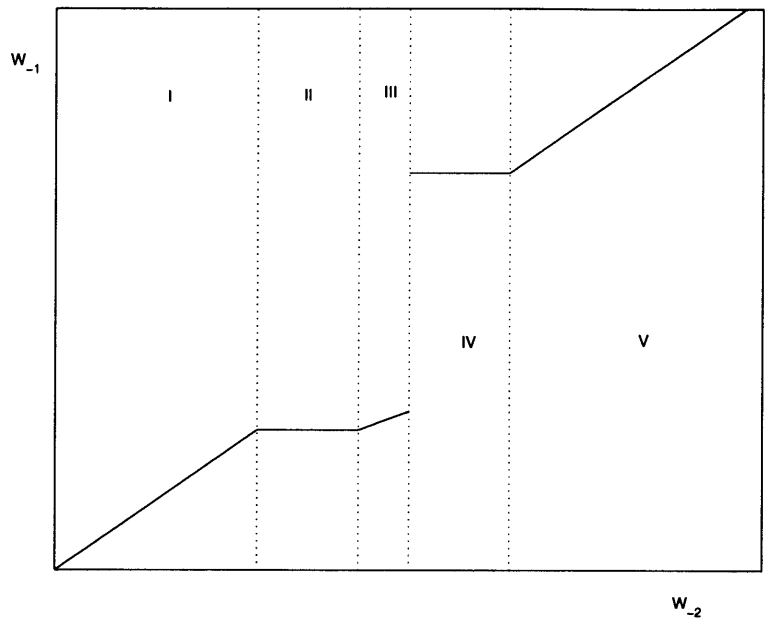

Figure 3-4: Savings of self -2

Note: the figure is only qualitative; it is not meant to illustrate actual slopes or relative sizes for the regions.

Region IV is the content of lemma 4–self -2 chooses early retirement, but in order for her to do that, she needs to oversave, otherwise self -1 ends up making self 0 work. Again, in this region self -2 consumes all extra marginal wealth, until she is rich enough so that eventual early retirement results without oversaving. And in region V, self -2 consumes according to the Laibson solution, leading to early retirement.

More generally, for any wealth level W_t with $t \leq -2$ such that self t wants self 0 to retire, self $t-1$ wants her to retire as well. The converse of this is not true, i.e. if self t chooses to save so that self 0 works, self $t-1$ might not like that. Translating

prove by backward induction and taking advantage of the above lemma that two things are possible. Either self t's marginal propensity to consume is λ^* up to $(\frac{1}{(1-\lambda^*)R})^t \overline{W}_0$, and (if she still prefers late retirement for higher wealth levels) then her marginal propensity to consume is 1 on some non-empty interval. Or self t's marginal propensity to consume is λ^* up to some lower wealth level, above which she prefers early retirement. This immediately implies two things. First, if mandated work is acceptable to all selves (in the sense that they prefer late retirement at their mandate wealth level), then the outcome of a mandate is an equilibrium even with choice. Second, if this is not the case (the mandate is not acceptable to all selves), then savings for retirement in a work equilibrium without a mandate is lower. Also, if all selves prefer lower saving for at least some wealth levels, then small enough amounts of lower saving will all be done by the first self alive.

our intuition from the work equilibrium, we might be led to think that–due to the elimination of this conflict–if retirement was mandated, savings levels would be lower. Such a conclusion is true in the present setup, but not if we go back one more period. Imagine that with the mandate, W_{-1} is slightly above $\overline{W}_{-1}$, the cutoff wealth level for self -1, and that self -2 is willing to bequeath higher savings to make self -1 choose early retirement. That is, even for some wealth levels below $\frac{1}{R}\frac{1}{1-\lambda^*}\overline{W}_{-1}$, self -2 will choose to save $\overline{W}_{-1}$. Since self -2 overconsumes from the point of view of self -3, self -3 might choose to lower her savings to self -2 once the mandate is removed. Then self -1 will end up with $\overline{W}_{-1}$–lower than with the mandate. The key intuition is that self -3 takes advantage of self -2's efforts to control self -1's decision for her own purposes[25].

This highlights a key distinction between the early and late retirement outcomes. When there is higher saving to be done to force retirement, the earlier selves are by no means as eager to join in as when the task is lower saving. They are actually very happy to let later selves save more, as those selves consume too much from their point of view anyway. They will thus want to have them oversave a lot, often resulting in putting the self at her cutoff wealth level. (In more precise language: a self t who is leaving boundary saving but over her own cutoff wealth level usually has a marginal propensity to consume of 1, thus making the marginal rate of substitution for self t-1 low.) As a consequence, small amounts of lower savings are 'handled' by the early selves, while higher saving is pushed on (in an exaggerated manner, in fact) to later ones.

Partly for this reason, it is important to focus on the lower saving outcome if we care about the well-being of the individual as a whole, that is, the set of her intertemporal incarnations. For such an analysis we can use similar tools as in welfare economics. In the 'forced' work outcome, the implications are generally bad. The Laibson consumption path is already too high, and there is additional consumption

[25]The same counterexample works to show that the other statement from the late retirement case does not carry over, either: it is not true that if mandated retirement is acceptable to all selves, then the outcome of the mandate is an equilibrium.

done in the periods before retirement, making the equilibrium outcome Pareto-inferior to a mandate: self -2 would benefit from a better consumption path, selves 0 and up from more savings, and self -1 (possibly) from both. In this strong sense, the equilibrium outcome is suboptimal, and can correctly be termed an undersaving outcome. Similarly unambiguous things cannot be said when the equilibrium has retirement in period 0. Higher saving by a self is in general good for both earlier and later selves but bad for that self. So, on the one hand, commitment might not be desirable, and on the other, its welfare implications are mixed.

In all these proofs we have *very strongly* used the particular structure of quasi-hyperbolic discounting[26]. A troublesome occurrence of this was when we proved that in periods $t \leq -2$ lower saving is not possible in the boundary savings level sense (lemma 4): the proof depended on the fact that selves t and $t-1$ have two different weightings of the same utility tradeoff (c_t vs. K_t). Since Laibson introduced quasi-hyperbolic discounting as an approximation to hyperbolic discounting purely for analytical convenience, such results should be handled with great suspicion. In addition, our intuition should revolt at strange answers of this sort. In a true hyperbolic discount structure, from the point of view of self -2, self -1 not only underweights effort in period 0 compared to consumption in period -1, but she also overweights it compared to consumption after retirement. This results in self -2 choosing to undersave more often than in a quasi-hyperbolic model, where the second conflict is nonexistent. For a formal treatment, see the appendix.

3.6 Notes on observational equivalence

One of the important caveats of quasi-hyperbolic discounting is that it is very hard to tell it apart from exponential discounting. Laibson (Laibson 1996) noted that an econometrician watching a quasi-hyperbolic discounter, but operating under the assumption of exponential discounting, will get a very good fit for her theory, as

[26] Even the appendix's proof of the existence of equilibria uses at a crucial point that with quasi-hyperbolic discounting all earlier selves would want a later self to do the same thing.

consumption paths of the two types of agents look exactly the same. At the same time, she will radically misconstrue the agent's preferences, finding a one-period discount factor of 0.98 instead of 0.6 in a typical example. Only comparative statics involving the interest rate can be used to distinguish actors with self-control problems from the others.

Our models lend themselves to a number of convenient approaches to this question. Both the consumption path and some comparative statics results can give a quasi-hyperbolic discounter away.

First, a consumption path that is smooth after retirement and not smooth leading up to it is a sign of quasi-hyperbolic discounting. This is of course due to the changing marginal propensities to consume in the periods preceding retirement. In particular, if equilibrium involves work in the period of decision, a lower average consumption rate after retirement than before is consistent with quasi-hyperbolic but not with exponential discounting[27].

Interesting comparisons of comparative statics nature also emerge. Consider an equilibrium in the three-period model in which the agent undersaves in the first period and works in the second. If earnings in the second period (Δ) increase, the period 1 self will save *more*: the extra earnings gives self 2 more incentive to work, lowering the amount of undersaving needed to induce work. Thus, self 1's marginal propensity to consume out of changes in future earning is negative. This could never happen with an exponential discounter.

The complete comparative statics for first-period savings with respect to Δ is illustrated in figure 3-5 for the three-period model. Savings with and without commitment are shown. For very low levels of Δ it is not worth working, so the agent just saves from her other wealth for retirement. These savings don't depend on Δ, as period 2 income is never realized. In the next region, self 1 would prefer self 2 to work if she could commit her to do it, but, without it, it is better to retire early. The

[27]It is tempting at first to try to use this as an explanation for the drop in consumption at retirement. There a number of problems, though: first, the drop in consumption occurs at period t=-1 the latest, i.e. *before* retirement. Also, the drop is much too general of a finding for this theory: it happens to almost all groups of people, irrespective of wealth or when they retire.

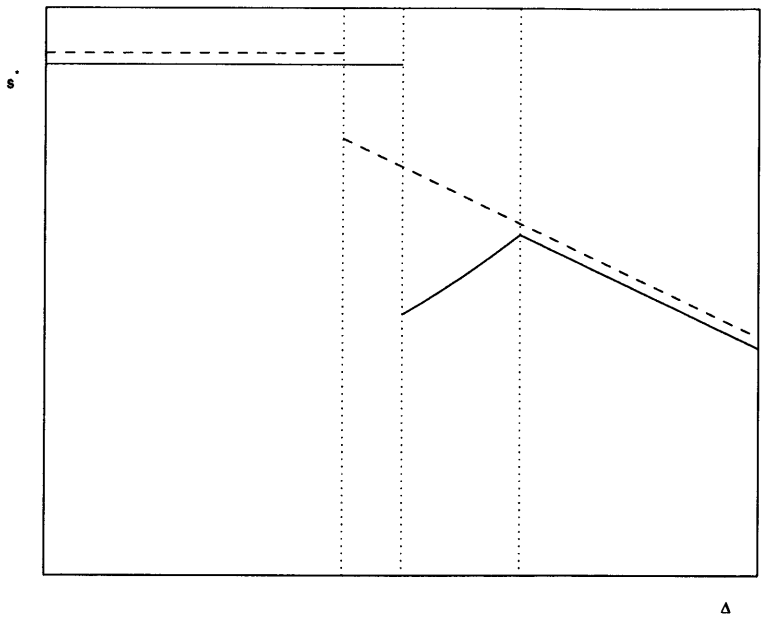

Figure 3-5: Savings of self 1 with (dashed line) and without (continuous line) commitment as a function of self 2's wage income
Note: the figure is only qualitative; it is not meant to illustrate actual slopes or relative sizes for the regions.

interesting region is the next one. Here, self 1 undersaves to make self 2 work, giving her exactly $s^* = \overline{W}_2$. Since $\overline{W}_2$ increases with Δ, s^* is increasing; furthermore, this is the only region in which s^* is in general not a linear function of Δ. And finally, for high levels of Δ the equilibrium involves work, and it is once again equivalent to the commitment solution. Notice that these regions are in exactly the opposite order as in figure 3-3–higher levels of wealth and higher levels of earnings have opposite incentive effects for retirement.

To check for this effect in practice, one might use the path of consumption to identify people who are undersaving, and then check whether agents with a higher income-to-wealth ratio save a larger proportion of their wealth for the period of decision.

Similarly, if the option to retire in the second period is eliminated in some way, and self 1 would have undersaved before, she will save more. This is again impossible with exponential discounting: there the elimination of a non-chosen alternative

doesn't change the optimum. Also, agents who work in the second period in equilibrium but don't undersave, will not change their behavior. Therefore, it is probably econometrically easier to check this effect: it is not necessary to identify undersavers, an exercise that carries with it the backbreaking task of disentangling undersaving from other peculiarities in the consumption path that happen around retirement. It is, however, necessary to identify those who would have worked had the option been available.

On the other hand, if the option to work is eliminated, the comparative statics are uncertain. We have seen that savings could go up or down in this case.

Finally, notice that in the long-horizon equilibrium there is a range of wealth levels where richer people save disproportionately more of their wealth for retirement: as one switches from lower savings and work to retirement, the total wealth that self 0 gets switches from $\overline{W}_0$ to something that is greater than $\overline{W}_0 + \lambda^*\Delta$, a change that is not warranted by the difference in lifetime wealth. Thus, controlling for income, on average the richer people (who retire early) have higher savings rates. This is 'anecdotally' consistent with the stylized fact that wealthier people appear to be slightly more patient, a finding that can't be explained by exponential discounting without appealing to individual heterogeneity in time preference.

3.7 Naive agents

As economists, we often assume too much rational capability on the part of humans. Our assumption of full sophistication of the agents is not immune from this criticism. Thus, we study less-than-sophisticated agents in this section. The opposite extreme assumption to sophistication is naivité.

An agent is called naive if each of her intertemporal selves assumes that future selves will make the same consumption and/or retirement decisions as she would. There is no game in this case, and the 'plans' (current decisions and expectations about future decisions) are simply updated each period. For a contrast of sophistication and naivité in the context of quasi-hyperbolic discounting, see (O'Donoghue and

Rabin 1999a).

We work in the long-horizon setting. First, let us assume that the agent is naive only about the retirement decision, not consumption. That is, she still plays a Laibson game with respect to consumption, but each self $t < 0$ assumes that self 0 (and others) will make the same retirement decision as she would. This assumption is mostly for analytical convenience, so that the discussion fits more naturally into what we have been doing. But it might also be interesting empirically, because retirement is (mostly) a one-time decision, so people should have less of a chance of learning about their intertemporal conflicts in this area than regarding consumption[28]. The assumption implies that self $-n < -1$ expects to work in period 0 iff

$$V(W) - \beta\delta^n e \geq V(W - \frac{1}{R^n}\Delta), \tag{3.16}$$

where period $-n$ wealth, W, now includes discounted earnings in period 0. We have the following result:

Theorem 14 *If self $-n < -1$ plans to work in period 0, so does self $-n+1$.*

This theorem is the consequence of two considerations, one specific to quasi-hyperbolic discounting and one not. First, the Euler equation for consumption implies that the marginal utility of wealth today is less than $R\delta$ times the marginal utility tomorrow, so an extra amount of income that is R times as much in the future as today should be worth more than $\frac{1}{\delta}$ times in the future as today. And since in the future the cost will be perceived to be $\frac{1}{\delta}$ times as much, the future self is more likely to want to work. Second, the future self is additionally motivated to want to work as the current self consumes part of the planned income in period 0. The latter argument does not rely on quasi-hyperbolic discounting, while the former one does. The proof of theorem 14 takes advantage of both in a tricky way. It could be simplified, but the given form allows for two generalizations.

[28]Note that with this assumption consumption decisions in each period are the same as in the commitment case.

It is easy to show that if $(\lambda^*\beta + 1 - \lambda^*)^n \leq \beta$, then self 0 will actually work. Also, though the problem seems different on the surface, both the theorem and the conclusions below are exactly the same if the agent is also naive about consumption decisions.

The converse of theorem 14 is not true–if self $-n < -1$ wants to retire early, self $-n + 1$ might change her mind.

It is, however, true that if self -1 wants to retire in period 0, self 0 will actually do so. To see this, note that if self 0 were to work, that would be better for self -1 as well, and with optimal consumption it would be better still. Again, the converse of this is not true: it could happen that self -1 plans to work in period 0, but self 0 decides to retire. In this case, lifetime wealth is updated downwards (self -1 believes that period 0 earnings are a part of wealth), so there is a downward jump in the consumption path. In contrast to the sophisticated case, this occurs *exactly* at retirement, as actually observed empirically (Bernheim, Skinner, and Weinberg 1997).

3.8 Conclusion

This paper makes an addition to the classic quasi-hyperbolic discounting savings model. Its technical contributions are minor–most of the analysis is possible with little more than the tools developed by David Laibson. However, the interaction of two decisions, with the one (savings) available as a tool to influence the other (retirement), changes the classic model in a few interesting ways.

One is the possibility of additional undersaving with the eventual consequence of making the self with a choice poor enough so that she will want to work. This undersaving occurs *in addition* to the undersaving that characterizes the equilibrium without a retirement decision. It therefore aggravates an already inefficient outcome, and, not surprisingly, is likely to be bad for all selves.

The other, and perhaps more novel, effect is the possibility of higher saving. Higher saving can occur for two reasons: either because it is too costly in terms of discounted utility to make the deciding self (self 2 in the three-period models, self

0 in the others) work, and thus one would rather finance her retirement, or because self $t \leq -1$ is too eager to work long and it is worth saving more to make her choose early retirement. Unlike undersaving, it is not in general bad for the individual–it can mitigate the overconsumption equilibrium of the classic model. In fact, higher saving seems never to be Pareto-worsening: the later selves, at least, should be happy about getting more savings. For this reason, it might not be as important in practice as the undersaving equilibrium.

Testing for quasi-hyperbolic discounting empirically runs into one major problem: individual marginal propensities to consume, which are crucial in these models, are basically impossible to measure. Thus a direct test of the Euler equation for consumption is practically infeasible. The retirement decision parameter, on the other hand, allows for good indirect tests, based on either the consumption path or comparative statics. It might be interesting to see some of these tests done. Particularly interesting would be 'natural experiments' changing work and retirement opportunities.

The theoretical model would benefit from two major extensions. One is the introduction of more periods when the agent can choose whether to work. We solved a model of this sort without savings: in each period, the agent can decide whether or not to retire (the retirement decision being final,) and consumption just equals income or benefits. To make it an interesting problem, one has to assume, for example, a benefit profile that increases with the age of retirement. In equilibrium, the agent retires too early: the retirement date is Pareto-dominated by a later retirement date. No such results emerge in our models with savings, but they might if there are more periods of retirement decisions[29]. Also, it would be interesting to see how consumption unfolds during the periods of decision, and how the different retirement dates are distributed among these periods.

The other useful extension would be the investigation of liquidity constraints in this context. They are clearly important in practice, and they change the nature of

[29]The Pareto-improving retirement date is at least two periods later than the equilibrium date t: otherwise self t wouldn't want to retire. Then it is not a major surprise that our models don't generate too early retirement.

equilibria with quasi-hyperbolic discounting considerably.

A perplexing aspect of quasi-hyperbolic discounting models is a question that is very hard to answer: why don't people take advantage of annuity-type commitment devices to overcome their undersaving problem? These financial tools are readily available but rarely used. Some modestly satisfactory reasons can be brought up. First, if there is a bequest motive, then, just like in many exponential discounting models, annuities look less attractive than without a bequest motive. Second, the annuities market is quite complicated, and there are good reasons for boundedly rational people not to enter markets they know little about. The latter seems to indicate that as people learn about annuities they will come into broader use. Even if that happens, the commitment is unlikely to be full, leaving at least some room for quasi-hyperbolic discounting effects. In the absence of annuities, there is of course a wide-spread institutional structure that serves as a commitment device for agents happy or unhappy about it: social security. We plan to study the implications of the joint mandates of savings and receipt of social security benefits as a real annuity in a later paper.

.1 Proof of Theorem 1

Theorem 1 *An optimal strategy exists.*

The following proof is very general, and works for a broad class of these problems. We only sketch the proof.

Proof. For periods 1 and 2, the statement is trivial. Also, expected utility is bounded and it is continuous in the beliefs held at the beginning of period 1 except at zero mean beliefs.

A stopping rule for period 0 specifies, for each posterior belief, whether to go on sampling information or not. The set of stopping rules is thus a product of two-point sets, which is compact in the product topology by Tychonoff's theorem (e.g. Pryce (1973)). The set of measurable stopping rules is a closed subset of this set, and thus is also compact.

A belief at the beginning of period 1 is a measure over $\Re$. A stopping rule induces a measure over these beliefs. Give the set of measures over beliefs the weak-* topology. Lebesgue's dominated convergence theorem (Stroock 1994) implies that the map between stopping rules and measures is continuous. This implies that the set of measures over beliefs available to the agent is compact. Finally, expected utility is continuous in this measure. □

.2 Other Proofs

Lemma 5 *Let current beliefs satisfy*

$$\begin{aligned} p_{0H} &= p_{1H} + \epsilon_H \\ p_{0L} &= p_{1L} + \epsilon_L \\ c &= Prob(q^t = q_H^t) = p_{1H} + p_{0H}. \end{aligned} \tag{17}$$

Then, a type-dependent signal about option 1 is uninformative about ability if and only if

$$(1-c)\epsilon_H(q_H^t - \frac{1}{2}) = c\epsilon_L(q_L^t - \frac{1}{2}). \tag{18}$$

Proof. By the law of iterated expectations, a necessary and sufficient condition for the beliefs about q^t not to move is to not have them move after the signal $s_1 = 1$. We have

$$\begin{aligned} Prob(q^t = q_H^t | s^1 = 1) &= \frac{p_{1H}q_H^t + p_{0H}(1-q_H^t)}{p_{1H}q_H^t + p_{0H}(1-q_H^t) + p_{1L}q_L^t + p_{0L}(1-q_L^t)} \\ &= \frac{p_{1H} + \epsilon_H(1-q_H^t)}{p_{1H} + \epsilon_H(1-q_H^t) + p_{1L} + \epsilon_l(1-q_L^t)} \\ &= \frac{\frac{1}{2}c + \epsilon_H(1-q_H^t - \frac{1}{2})}{\frac{1}{2} + \epsilon_H(1-q_H^t - \frac{1}{2}) + \epsilon_L(1-q_L^t - \frac{1}{2})} \end{aligned} \tag{19}$$

Setting this equal to c gives the result. □

For $\epsilon_H = \epsilon_L = 0$, the result should be clear: if a_2 is independent of q and $a_2 = 0$ is just as likely as $a_2 = 1$, then no signal is informative about ability–in essence, the signal is the 'first' piece of information about the choice, and can't confirm or disconfirm previously held beliefs. It is, however, somewhat surprising that even if a_2 is not independent of q (ϵ_H, ϵ_L non-zero), the signal might not be informative. To understand this, consider $c = \frac{1}{2}$ and $\epsilon_H > 0$. One might think that $s^1 = 1$ is bad news: since it is more likely that $a_2 = 0$, chances are the signal is wrong, or that the agent is of low type. But if ϵ_L is sufficiently large, one wouldn't expect $s^1 = 1$ from

low types, either, so the signal doesn't tilt beliefs in the negative direction.

Lemma 6 *Using the notation of lemma 5, a type-independent signal is uninformative about ability if and only if*

$$(1-c)\epsilon_H = c\epsilon_L. \tag{20}$$

Proof. Similar to that of lemma 5.

Lemma 1 *A type-dependent signal followed by any other signal is always informative about ability.*

Proof. If the first signal is informative, then so are the two of them together– beliefs after the two of them are just a mean-preserving spread of the beliefs after the first. Thus, it is sufficient to prove that if the first signal is not informative, then the second one is.

To prove this for two type-dependent signal, first note that the condition in lemma 5 can only hold for both signals if $\epsilon_H = \epsilon_L = 0$ or the ratio $\frac{q_H - \frac{1}{2}}{q_L - \frac{1}{2}}$ is the same across the time periods. So unless this is the case, one of the signals is already informative, and order doesn't matter.

To complete the proof, we prove if one of the above two conditions holds, the relationship given in equation 18 can't be preserved after updating. The ratio corresponding to the ratio of ϵ_H and ϵ_L after updating is

$$\frac{p_{0H}(1-q_H^t) - p_{1H}q_H^t}{p_{0L}(1-q_L^t) - p_{1L}q_L^t} = \frac{\frac{1}{2}\epsilon_H - (q_H^t - \frac{1}{2})c}{\frac{1}{2}\epsilon_L - (q_L^t - \frac{1}{2})(1-c)}. \tag{21}$$

If $\epsilon_H = \epsilon_L = 0$, a trivial use of lemma 5 shows that the second signal is informative. In the other case, the above ratio should be equal to $\frac{\epsilon_H}{\epsilon_L}$. For this we would need to have

$$\frac{\epsilon_H}{\epsilon_L} = \frac{c}{1-c}\frac{q_H^t - \frac{1}{2}}{q_L^t - \frac{1}{2}}, \tag{22}$$

which, by lemma 5, is not true if the first signal is uninformative.

Now to prove that a type-dependent followed by a type-independent signal is informative, notice that the conditions of lemmas 5 and 6 can only hold at the same time if $\epsilon_H = \epsilon_L = 0$. So unless this is the case, we are done–otherwise, once again, one of the signals is informative by itself. And even if $\epsilon_H = \epsilon_L = 0$, the above proof shows that after updating using the period 1 signal, the posteriors won't satisfy the conditions of lemma 6. □

The proof takes advantage of the fact that in order for the first signal to be uninformative, the absolute value of ϵ_H has to be smaller than the absolute value of ϵ_L. But conditional on being type H, a signal is more informative, so ϵ_H is going to be moved by more than ϵ_L. Thus the ratio can't be preserved.

.3 Existence and uniqueness of equilibria

In this section, we outline a proof of the existence of equilibrium for the long-horizon game. It just requires pulling together much of what we have already shown.

For the game after retirement, the existence and uniqueness of the subgame-perfect equilibrium has been established by David Laibson. For earlier periods, we prove the following general theorem:

Lemma 7 *A Markov-perfect subgame-perfect equilibrium exists with the following properties. For $t \leq 0$, the domain $(0, \infty)$ of the consumption rule $c_t(W_t)$ can be divided into finitely many disjoint intervals such that in the interior of each interval,*

1. the eventual period 0 work/retirement decision is the same,

2. the equilibrium consumption schedules $c_s(W_t)$ for $s > t$ are all differentiable in W_t, and

3. self t has a constant marginal propensity to consume;

4. further, at an interval endpoint a, self t is indifferent between following the limit of the two neighboring intervals' consumption rules, and utility is continuous in wealth at a.

Proof. Starting from $t = 0$, use the following backward induction type of construction for finding the equilibrium: given the next self's strategy, maximize utility for self t. If for some wealth self t is indifferent between a number of consumption levels, assign to her the strategy that the earlier self would prefer.

Of course, we have to prove that this construction works and yields an equilibrium with the above properties. We do this by backward induction.

The case is clear for $t = 0$. Now suppose the statement is true for $t = m + 1$. We will prove it for $t = m$.

Suppose W_m is given. For self $m + 1$, let the intervals in question be divided the by points $0 < a_1 < \ldots < a_M$. For any $\epsilon > 0$, self m's maximization problem has a solution if her savings level is restricted to lie in the interval $[a_i + \epsilon, a_{i+1} - \epsilon]$. Since there are only finitely many intervals, a maximum on the union of these intervals and the points $\{a_i\}$ also exists. It is easy to see that as ϵ approaches zero, eventually

the maximum doesn't change. For otherwise there would be a point a_i such that as W_{m+1} approaches a_i from one of the sides, self m's utility is greater than at savings level a_i, which contradicts that when indifferent, self $m+1$ chooses the consumption level self m prefers[30].

This shows that for each wealth level W_m, self m's problem has a solution. Now define $0 = b_0 < b_1 < \ldots < b_N < \infty$ such that for each $i = 0, \ldots, N-1$, if $W_m \in (b_{2i}, b_{2i+1})$, then $W_{m+1} \in (a_i, ai+1$, and if $W_m \in (b_{2i+1}, b_{2i+2})$, then $W_{m+1} = a_{i+1}$[31].

By definition, point 1 is satisfied for each (b_j, b_{j+1}). It is also clear that for any (b_{2i+1}, b_{2i+2}), points 2 and 3 are satisfied as well. Therefore let us concentrate on the case $W_m \in (b_{2i}, b_{2i+1})$. Since all future consumption schedules are differentiable at $W_{m+1}(W_m)$, the discounted utility of self m as a function of c_m is differentiable at $c_m(W_m)$. Now $c_m(W_m)$ maximizes this utility, so the derivative at that point is zero. Taking the derivative for selves m and $m+1$ (as in Laibson (Laibson 1996)) and substituting leads to the Euler equation

$$\frac{u'(c_m)}{u'(c_{m+1})} = R\delta(\beta\lambda_{m+1} + 1 - \lambda_{m+1}), \tag{23}$$

where λ_{m+1} is self $m+1$'s marginal propensity to consume . Then self m's marginal propensity to consume λ_m on (b_{2i+1}, b_{2i+2}) is constant and is given by the equation

$$\frac{\lambda_m}{1-\lambda_m} = \frac{R\lambda_{m+1}}{[R\delta(\beta\lambda_{m+1} + 1 - \lambda_{m+1})]^{\frac{1}{\rho}}}. \tag{24}$$

(This is just Laibson's recursion for the λs.)

Also, clearly, utility is continuous in wealth at each interval endpoint, otherwise the agent would 'jump' to the other interval at a different place. Finally, we need to show the agent is indifferent between the limits of the two neighboring consumption

[30]More precisely, there is a sequence $W_{m+1,n}$ approaching a_i from one side such that discounted utility for self m is increasing on that sequence, and the limit of the discounted utilities is more than discounted utility at a_i. But if at wealth level a_i self $m+1$ consumes $\lim c_{m+1}(W_{m+1,n})$, by point 4 the discounted utility of self m should be the limit of the discounted utilities when leaving savings $W_{m+1,n}$. But this is impossible by construction as we have assumed that when indifferent, self $m+1$ does what self m prefers.

[31]Of course, some of the intervals (b_j, b_{j+1}) may be empty.

rules. Suppose by contradiction that, say, consuming $\lim_{W_t \searrow a} c_t(W_t)$, doesn't yield the limit of the utilities. This could only be because one of the future selves jumped at an interval endpoint. Then self m's utility actually increased, because when indifferent future selves do what self m wants them to (with quasi-hyperbolic discounting, all previous selves want the same thing.) But in this case near a self m's choice of consumption wasn't optimal, a contradiction. □

In terms of outcome, and with the qualification that the first self alive might be indifferent between two consumption levels at finitely many points, this equilibrium seems to be unique up to outcome: if an interval endpoint is a maximum for an earlier self for some wealth level, then the assumption that the self does what the earlier one prefers is necessary (otherwise that self's problem has no solution), and else it is irrelevant.

.4 Proofs of some claims

To prove lemma 3, we need the following preliminary result.

Lemma 8 *For $t \leq -1$, savings is monotonically increasing in wealth.*

Proof. Suppose by contradiction that $W_t > W_t'$ but that the corresponding savings levels satisfy $W_{t+1} < W_{t+1}'$. Let the consumption levels be c_t and c_t' and denote the continuation utilities from leaving wealth levels W_{t+1} and W_{t+1}' by K and K', respectively. Further, define $c_t'' = W_t - \frac{1}{R}W_{t+1}'$, $c_t''' = W_t' - \frac{1}{R}W_{t+1}$. Then

$$u(c_t'') + K' \leq u(c_t) + K \tag{25}$$

$$u(c_t''') + K \leq u(c_t') + K'. \tag{26}$$

We can add these and eliminate K and K' to get

$$u(c_t'') + u(c_t''') \leq u(c_t) + u(c_t'). \tag{27}$$

But notice that $c_t > c_t'', c_t''' > c_t'$ and $c_t + c_t' = c_t'' + c_t'''$. Since u is concave, the inequality 27 is impossible. This completes the proof. □

Lemma 3 *Suppose that $t \leq 0$ and $W_t > W_t'$. Then it is not possible that self t with wealth W_t behaves so that self 0 eventually works, and with wealth W_t' she behaves so that self 0 eventually retires.*

Proof. We prove by backward induction. The statement is clearly true for $t = 0$.

Suppose the statement is true for $t = m + 1$. We will prove by contradiction that it is true for $t = m$. Suppose it isn't. Then there are wealth levels W_m and W_m' such that $W_m > W_m'$ and with wealth W_m self 0 eventually works, and with wealth W_m' self 0 eventually retires. Since our statement is true for $t = m + 1$, we then need to have $W_{m+1} < W_{m+1}'$. But this is impossible by lemma 8. □

Lemma 4 *Let $\overline{W}_t$ $(t < 0)$ be the level of wealth at which self t is indifferent between behaviors that eventually lead to self 0 working or retiring*[32]*. At this savings level, self $t - 1$ strictly prefers self t to choose to eventually make self 0 retire.*

Proof. We again prove by backward induction, although, as the reader will see, the need for that is little more than technical. Let c_t^w, K_t^w and c_t^r, K_t^r be the consumption levels and continuation utilities for self t with wealth level $\overline{W}_t$ in the working and retirement cases, respectively.

Suppose first that $t = -1$. We have $c_{-1}^w > c_{-1}^r$, since otherwise self -1 would have to leave $\overline{W}_0$ for self 0, which would not make him indifferent between working and retiring. Also

$$u(c_{-1}^r) + \beta\delta K_{-1}^r = u(c_{-1}^w) + \beta\delta K_{-1}^w. \tag{28}$$

[32] *Though this fact is not necessary here, it should be said that $\overline{W}_t$ exists and is unique. That it exists can be seen from the consideration that both the set of savings levels where eventual early retirement is (weakly) preferred and where eventual late retirement is preferred are closed. This can be proven easily using backward induction. That it is unique follows from a variant of lemma 3 (the proof of which didn't use strict preferences) along with backward induction.*

To see what self -2 would want, we have to compare $\beta\delta u(c^r_{-1})+\beta\delta^2 K^r_{-1}$ and $\beta\delta u(c^w_{-1})+\beta\delta^2 K^w_{-1}$. This is easy:

$$\beta\delta u(c^r_{-1}) + \beta\delta^2 K^r_{-1} - \beta\delta u(c^w_{-1}) - \beta\delta^2 K^w_{-1} =$$
$$\beta\delta(u(c^r_{-1}) - u(c^w_{-1})) + \beta\delta^2(K^r_{-1} - K^w_{-1}) =$$
$$\beta\delta(u(c^r_{-1}) - u(c^w_{-1})) + \delta(u(c^w_{-1}) - u(c^r_{-1})) = (\delta - \beta\delta)(u(c^w_{-1}) - u(c^r_{-1})) > 0$$

If the statement is true for t=m+1, then since self m+1 is not indifferent between self m+2 working and retiring at $\overline{W}_{m+2}$, we have $c^w_{m+1} > c^r_{m+1}$. Then the same proof as above works. □

Theorem 14 *If self $-n < -1$ plans to work in period 0, so does self $-n+1$.*

Proof. If self $-n$ plans to work, next period's wealth is $(1-\lambda^*)RW$. Now

$$\left[V((1-\lambda^*)RW) - V\left((1-\lambda^*)RW - \frac{1}{R^{n-1}}\Delta\right)\right]\delta >$$
$$> \left[V((1-\lambda^*)RW) - V\left((1-\lambda^*)RW - \frac{1}{R^{n-1}}\Delta\right)\right]\delta(\lambda^*\beta + 1 - \lambda^*) \qquad (29)$$

since $\beta < 1$. From equation 35 in the appendix, this equals

$$\frac{1}{R}\left(\frac{1}{1-\lambda^*}\frac{1}{R}\right)^{-\rho}\left[V((1-\lambda^*)RW) - V\left((1-\lambda^*)RW - \frac{1}{R^{n-1}}\Delta\right)\right] =$$
$$\frac{1}{R}\left(\frac{1}{1-\lambda^*}\frac{1}{R}\right)^{-\rho}\int_0^{\frac{1}{R^{n-1}}\Delta} V'((1-\lambda^*)RW - x)dx \qquad (30)$$

Since V is concave, the above is greater than

$$\frac{1}{R}\left(\frac{1}{1-\lambda^*}\frac{1}{R}\right)^{-\rho}\int_0^{\frac{1}{R^{n-1}}\Delta} V'((1-\lambda^*)RW - (1-\lambda^*)x)dx, \qquad (31)$$

which, since $V(W) = Vu(W)$ for each W, equals

$$\frac{1}{R}\int_0^{\frac{1}{R^{n-1}}\Delta} V'\left(W - \frac{1}{R}x\right)dx = \int_0^{\frac{1}{R^n}\Delta} V'(W-x)dx = V(W) - V\left(W - \frac{1}{R^n}\Delta\right) \qquad (32)$$

through a change in variables. Since self $-n$ plans to work, this is greater than or equal to $\beta\delta^n e$. But then $V((1-\lambda^*)RW)-V((1-\lambda^*)RW-\frac{1}{R^{n-1}}\Delta) \geq \beta\delta^{n-1}e$, implying the claim. □

The proof of theorem 14 really only used that

$$\frac{((1-\lambda^*)R)^\rho}{R\delta} < 1, \tag{33}$$

where λ^* is each self's marginal propensity to consume. Even for agents naive about consumption decisions, marginal propensity to consume is equal across periods with a value of

$$\lambda^* = \frac{1-(\delta R^{1-\rho})^{\frac{1}{\rho}}}{1-(1-\beta^{\frac{1}{\rho}})(\delta R^{1-\rho})^{\frac{1}{\rho}}}. \tag{34}$$

Assuming $\delta R^{1-\rho} < 1$, which is necessary for the naive maximization problem to have a solution, it is easily established that the above satisfies inequality 33. The proofs of the other claims in the text carry over quite effortlessly as well.

.5 A long horizon after retirement

To start, it is necessary to quote one of Laibson's results:

Lemma 9 *As the length of the horizon approaches infinity, the consumption rule converges pointwise to the function $c_t(W_t) = \lambda^* W_t$, where λ^* is the (unique) solution to the non-linear equation*

$$\lambda^* = 1-(\delta R^{1-\rho})^{\frac{1}{\rho}}[\lambda^*(\beta-1)+1]^{\frac{1}{\rho}}. \tag{35}$$

See (Laibson 1996), page 11 for this. Then we can define

$$\begin{aligned} V(W) &= u(\lambda^* W) + \beta\sum_{i=1}^{\infty}\delta^i u(\lambda^* R^i(1-\lambda^*)^i W) = V\frac{W^{1-\rho}}{1-\rho} = Vu(W), \text{ and} \\ D(W) &= \sum_{i=0}^{\infty}\delta^i u(\lambda^* R^i(1-\lambda^*)^i W) = D\frac{W^{1-\rho}}{1-\rho} = Du(W). \end{aligned} \tag{36}$$

These are the hyperbolically and exponentially discounted values of having wealth W[33]. For future reference, note that the constants V and D satisfy $\beta D < V < D$.

Suppose self 0 can make a retirement decision, but that selves before have to work and selves later have to retire. Self 0 will work iff

$$V(W_0 + \Delta) - e \geq V(W_0). \tag{37}$$

Let $\overline{W}_0$ be the wealth level that satisfies the above with equality. Self -1 will want self 0 to work iff

$$\beta\delta D(W_0 + \Delta) - \beta\delta e \geq \beta\delta D(W_0),$$

or

$$D(W_0 + \Delta) - e \geq D(W_0). \tag{38}$$

Since $D > V$, the inequality 38 will be satisfied more often than 37.

Analogously to section 3.3, self -1 would like to set $s_r = R(1 - \lambda^*)W_{-1}$ if self 0 doesn't work, and $s_w = R(1-\lambda^*)W_{-1} - \lambda^*\Delta$ if self 0 works. Now if the better of these options is available (doesn't conflict with self 2's preferred choice), that will be chosen. But it is also possible that self -1 prefers to work ($V(W_{-1} + \frac{1}{R}\Delta) - \beta\delta e > V(W_{-1})$, i.e. $s_c^* = s_w$), but $s_w = R(1 - \lambda^*)W_{-1} - \lambda^*\Delta$ is above $\overline{W}_0$. If the difference is not too large, it is preferable for self -1 to save less than s_c^* and make self 0 work. On the other hand, if the necessary reduction in savings to make self 0 work is a lot, self -1 will rather have self 0 retire with savings s_r.

These are just the cases we have seen before[34]. To see that $s^* < s_c^*$ can actually happen in 'practice,' revisit the hypothetical curves that we have seen in section 3.3. In the long-horizon case,

$$U_r(W_0) = \beta\delta D(W_0), \text{ and}$$

[33]Once again, for $u(c) = \ln(c)$, the constants V and D should enter additively. But also, the discussion is altered only trivially by this.

[34]And in all cases except when $s^* < s_c^*$, the consumption path looks like that in a simple savings game.

$$U_w(W_0) = \beta\delta D(W_0 + \Delta) - \beta\delta e, \tag{39}$$

and, as we have already seen, the optimal savings levels on the two curves are $s_w = R(1-\lambda^*)W_{-1} - \lambda^*\Delta$ and $s_r = R(1-\lambda^*)W_{-1}$. We know that $\overline{W}_0$ is defined by

$$V(\overline{W}_0 + \Delta) - V(\overline{W}_0) = e,$$

so define $\tilde{W}_0$ by

$$D(\tilde{W}_0 + \Delta) - D(\tilde{W}_0) = e.$$

We can easily choose the parameters such that $\tilde{W}_0 > \overline{W}_0 + \lambda^*\Delta$ (choose $\overline{W}_0$ and Δ such that $D(\overline{W}_0+\lambda^*\Delta+\Delta) - D(\overline{W}_0+\lambda^*\Delta) > V(\overline{W}_0+\Delta) - V(\overline{W}_0)$.) Then if W_{-1} is chosen such that s_w is just above $\overline{W}_0$, s_r will be less than $\tilde{W}_0$. Also, $U_w(W_0) > U_r(W_0)$ for $W_0 < \tilde{W}_0$, so even at s_r, the U_w curve dominates U_r. But since s_w is the optimal point on U_w, it clearly dominates the retirement alternative. Finally, if s_w is close enough to $\overline{W}_0$, the loss of utility from lower saving is small, so it will be worth doing it.

.6 A more hyperbolic discount structure

The only change we make is to introduce an additional discount parameter $\gamma < 1$ into Laibson's model, which is effective for 2 periods. Thus, self t's discounted utility from consumption is

$$u(c_t) + \beta\gamma\delta u(c_{t+1}) + \beta\gamma^2\delta^2 \sum_{i=0}^{\infty} \delta^i u(c_{t+2+i}). \tag{40}$$

Of course, we have to start from ground zero and solve the savings equilibrium before we can get into questions concerning retirement. The analysis is similar to Laibson's (Laibson 1996), and we will only go through an accelerated version of it.

Backwards induction along with a repeated use of property 3.2 of CRRA utility functions proves that in each period, consumption is a linear (and thus differentiable)

function of wealth. Then in equilibrium self t will choose c_t to satisfy

$$u'(c_t) = \beta\gamma\delta R\frac{\partial c_{t+1}}{\partial W_{t+1}}u'(c_{t+1}) + \beta\gamma^2\delta^2R^2\sum_{i=0}^{\infty}R^i\delta^i\frac{\partial c_{t+2+i}}{\partial W_{t+2+i}}\prod_{j=0}^{i}(1-\frac{\partial c_{t+1+j}}{\partial W_{t+1+j}})u'(c_{t+2+i}). \tag{41}$$

The similar equation for period t+1 is

$$u'(c_{t+1}) = \beta\gamma\delta R\frac{\partial c_{t+2}}{\partial W_{t+2}}u'(c_{t+2})+\beta\gamma^2\delta^2R^2\sum_{i=0}^{\infty}R^i\delta^i\frac{\partial c_{t+3+i}}{\partial W_{t+3+i}}\prod_{j=0}^{i}(1-\frac{\partial c_{t+2+j}}{\partial W_{t+2+j}})u'(c_{t+3+i}). \tag{42}$$

Combining the two we get

$$u'(c_t) = \beta\gamma\delta R\frac{\partial c_{t+1}}{\partial W_{t+1}}u'(c_{t+1}) + \beta\gamma^2\delta^2R^2\frac{\partial c_{t+2}}{\partial W_{t+2}}(1-\frac{\partial c_{t+1}}{\partial W_{t+1}})u'(c_{t+2}) + \delta R(1-\frac{\partial c_{t+1}}{\partial W_{t+1}})(u'(c_{t+1})-\beta\gamma\delta R\frac{\partial c_{t+2}}{\partial W_{t+2}}u'(c_{t+2})).$$

Putting this into a more convenient form leads to the following lemma.

Lemma 10 *The Euler equation for the choice of consumption at time t is*

$$u'(c_t) = \delta R\left(\beta\gamma\frac{\partial c_{t+1}}{\partial W_{t+1}} + (1-\frac{\partial c_{t+1}}{\partial W_{t+1}})\right)u'(c_{t+1})-\beta\gamma\delta^2R^2\frac{\partial c_{t+2}}{\partial W_{t+2}}(1-\frac{\partial c_{t+1}}{\partial W_{t+1}})u'(c_{t+2})(1-\gamma). \tag{43}$$

It is easily seen that for $\gamma = 1$ this reduces to Laibson's Euler equation.

Using this Euler equation, we can show that in a game with horizon T, the consumption rule is $c_t = \lambda_{T-t}W_t$, where the λ's are determined by the recursion

$$\left(\frac{\lambda_{n+2}}{1-\lambda_{n+2}}\right)^{-\rho} = \delta R^{1-\rho}(\beta\gamma\lambda_{n+1}+(1-\lambda_{n+1}))\lambda_{n+1}^{-\rho}-\beta\gamma\delta^2R^{2(1-\rho)}(1-\gamma)\lambda_n^{1-\rho}(1-\lambda_{n+1})^{1-\rho} \tag{44}$$

with initial value $\lambda_0 = 1$. Though we haven't shown that this converges, it seems to do so: in computer simulations it converged for all values of the parameters that we have tried[35]. The existence of a constant marginal propensity to consume far from the end is not technically necessary for what we are going to do, but it is nice to

[35]It must be said, though, that we haven't tried very many values.

work off a benchmark that has smooth consumption. We will therefore assume that for our parameter values the long-horizon case has a constant marginal propensity to consume of λ^*[36].

As before, we introduce variously discounted value functions. We will need three this time:

$$\begin{aligned} V(W) &= u(\lambda^* W) + \beta\gamma\delta u(\lambda^*(1-\lambda^*)RW) + \beta\gamma^2\delta^2 \sum_{i=2}^{\infty} \delta^{i-2} u(\lambda^* R^i (1-\lambda^*)^i W) \\ Z(W) &= u(\lambda^* W) + \gamma\delta \sum_{i=1}^{\infty} \delta^{i-1} u(\lambda^* R^i (1-\lambda^*)^i W) \\ D(W) &= \sum_{i=0}^{\infty} \delta^i u(\lambda^* R^i (1-\lambda^*)^i W) \end{aligned} \tag{45}$$

It is easy to see that one period before retirement we get the same undersaving possibility as with quasi-hyperbolic discounting. In that case $\overline{W}_{-1}$ is defined by

$$u(\overline{W}_{-1} - \frac{1}{R}\overline{W}_0) + \beta\gamma\delta Z(\overline{W}_0 + \Delta) - \beta\gamma\delta e = V(\overline{W}_{-1}). \tag{46}$$

To see what self -2 wants self -1 to do at this wealth level we want to look at the difference

$$\beta\gamma\delta u(\overline{W}_{-1} - \frac{1}{R}\overline{W}_0) + \beta\gamma^2\delta^2 D(\overline{W}_0 + \Delta) - \beta\gamma^2\delta^2 e - \beta\gamma\delta Z(\overline{W}_{-1}). \tag{47}$$

Using that $V(W) = \beta\gamma Z(W) + (1-\beta\gamma)u(\lambda^* W) + \beta\gamma\delta(1-\beta\gamma)u(\lambda^*(1-\lambda^*)RW)$ and $Z(W) = \gamma D(W) + (1-\gamma)u(\lambda^* W)$, along with equation 46, the above becomes

$$\begin{aligned} &\beta\gamma\delta u(c_{-1}^w) + \beta\gamma^2\delta^2 D(\overline{W}_0 + \Delta) - \beta\gamma^2\delta^2 e - \beta\gamma^2\delta^2 D(\overline{W}_0 + \Delta) - \\ &\beta\gamma\delta^2(1-\gamma)u(c_0^w) + \beta\gamma\delta^2 e + \delta(1-\beta\gamma)u(c_{-1}^r) + \beta\gamma\delta^2(1-\gamma)u(c_0^r), \end{aligned} \tag{48}$$

where the subscripts on c denote the period in question and the superscripts stand

[36] In this case, we also get the familiar undersaving outcome.

for whether retirement or work is chosen. Dividing by δ and regrouping we get

$$-(1-\gamma)[(u(c^w_{-1}+\beta\gamma\delta u(c^w_o)-\beta\gamma\delta e)-(u(c^r_{-1})+\beta\gamma\delta u(c^r_0))]-\gamma(1-\beta)(u(c^w_{-1})-u(c^r_{-1})). \tag{49}$$

Using that self -1 is indifferent between working and retiring

$$(1-\gamma)[\beta\gamma^2\delta^2 D((1-\lambda^*)R(\overline{W}_0+\Delta))-\beta\gamma^2\delta^2 D((1-\lambda^*)^2R^2\overline{W}_{-1})]-\gamma(1-\beta)(u(c^w_{-1})-u(c^r_{-1})). \tag{50}$$

Dividing by γ, we finally get that the difference 47 has the same sign as

$$\beta\gamma\delta^2(1-\gamma)[\overbrace{D((1-\lambda^*)R(\overline{W}_0+\Delta))-D((1-\lambda^*)^2R^2\overline{W}_{-1})}^{I}]-(1-\beta)(\overbrace{u(c^w_{-1})-u(c^r_{-1})}^{II}). \tag{51}$$

The second term in this sum is always negative (II is positive), while the other one can be either positive or negative, though it seems it is more often positive (for that we only need $\overline{W}_0+\Delta > R(1-\lambda^*)\overline{W}_{-1}$). For $\gamma = 1$, the first term drops out, so the expression is negative, which means that self -2 would want self -1 to retire at this wealth level. This is just what we had before. On the other hand, with $\gamma \neq 1$ and no degeneracy, the first term can be positive, so we don't necessarily get a negative sum. In particular, if the first term is positive and $\beta = 1$, we can only get lower saving (that is, self -2 wants self -1 to work at $\overline{W}_{-1}$.)

In general, both for γ and $1-\gamma$ close to 0 (both relative to $1-\beta$), we will get higher saving. This will be clear intuitively as soon as we understand that equation 51 contrasts two conflicts between selves -1 and -2. First, from the point of view of self -2, self -1 discounts too much between periods -1 and 0, as we had before (term II). But also, self -1 discounts too much between periods 0 and 1, that is, she doesn't appreciate the extra consumption from working as much as she should (term I). For β close to 1, the first effect is negligible. For γ close to 0 or 1, the second one is: close to 1 because then the conflict is small, and close to 0 because then the effect is 'too far in the future' (it is very discounted).

This is only a simple extension of the quasi-hyperbolic setup, but it still indicates

that lower saving is more likely with hyperbolic discounting. It also captures what appears to be the two most important conflicts between selves -1 and -2 regarding retirement: that from the perspective of self -2, self -1 overweights consumption in period -1 but underweights consumption after period 0 relative to effort in period 0. Their conflicts about consumption in periods after period 0 are likely to be unimportant. Of course, for earlier selves, this discount structure might not be sufficient: it would be interesting to see better approximations. It won't be easy: genuine hyperbolic discount functions generate equilibria that are extremely hard to analyze.

Bibliography

AINSLIE, G. (1975): "Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control," *Psychological Bulletin*, pp. 463–96.

AINSLIE, G., AND N. HASLAM (1992): "Hyperbolic Discounting," in *Choice Over Time*, ed. by G. Loewenstein, and J. Elster, p. 59. Russell Sage Foundation, New York.

AKERLOF, G., AND W. T. DICKENS (1982): "The Economic Consequences of Cognitive Dissonance," *The American Economic Review*, 72(3), 307–319.

ALTONJI, J. G., F. HAYASHI, AND L. J. KOTLIKOFF (1992): "Is the Extended Family Altruistically Linked? Direct Tests Using Micro Data," *American Economic Review*, 82(5), 1177–1198.

ARKIN, R. M., AND A. H. BAUMGARDNER (1985): "Self-Handicapping," in *Attribution: Basic Issues and Applications*, ed. by J. H. Harvey, and G. Weary, pp. 169–202. New York Academic Press.

BAKER, G. P., M. C. JENSEN, AND K. J. MURPHY (1988): "Compensation and Incentives: Practice vs. Theory," *The Journal of Finance*, 43(3), 593–616.

BARON, J. (1993): *Morality and rational choice.* Dordrecht ; Boston : Kluwer Academic Publishers.

BAUMEISTER, R. F. (1998): "The Self," in *The Handbook of Social Psychology*, ed. by D. Gilbert, S. Fiske, and G. Lindzey, pp. 680–740. Boston: McGraw-Hill.

BENABOU, R., AND J. TIROLE (1999a): "Self-confidence: Interpersonal Strategies," Mimeo.

——— (1999b): "Self-confidence: Intrapersonal Strategies," Mimeo.

BERNHEIM, D. B. (1994): "Do Households Appreciate their Financial Vulnerabilities? An Analysis of Actions, Perceptions, and Public Policy," *Mimeo, Princeton University.*

BERNHEIM, D. B., J. SKINNER, AND S. WEINBERG (1997): "What Accounts for the Variation in Retirement Wealth Among U.S. Households?," *Working Paper.*

BRANDENBURGER, A. M., AND B. J. NALEBUFF (1997): *Co-opetition: 1. A Revolutionary Mindset that Redefines Competition and Cooperation; 2. The Game Theory Strategy that's Changing the Game of Business.* Currency Doubleday.

BROCK, T. C., AND J. C. BALLOUN (1967): "Behavioral Receptivity to Dissonant Information," *Journal of Personality and Social Psychology*, 6, 413–428.

BROWN, J. D. (1986): "Evaluations of Self and Others: Self-enhancement Biases in Social Judgments," *Social Cognition*, 4, 353–376.

BURGER, J. M. (1986): "Temporal Effects on Attributions: Actor and Observer Differences," *Social Cognition*, 4(4), 377–387.

BURGER, J. M., AND R. M. HUNTZINGER (1985): "Temporal Effects on Attributions for One's Own Behavior: The Role of Task Outcome," *Journal of Experimental Social Psychology*, (21), 247–261.

CAMPBELL, J. D. (1986): "Similarity and Uniqueness: The Effects of Attribute Type, Relevance, and Individual Differences in Self-Esteem and Depression," *Journal of Personality and Social Psychology*, 50(2), 281–294.

CAPLIN, A., AND J. LEAHY (1999): "The Supply of Information by a Concerned Expert," C.V. Starr Center for Applied Economics Working Paper.

CARILLO, J. (1997): "Self-control, Moderate Consumption, and Craving," Mimeo.

CARILLO, J., AND T. MARIOTTI (1997): "Wishful Thinking and Strategic Ignorance," Mimeo.

COOPER, J., AND R. H. FAZIO (1984): "A New Look at Dissonance Theory," in *Advances in Experimental Social Psychology*, vol. 17. Academic Press.

COOPER, J., AND S. WORCHEL (1970): "Role of Undesired Consequences in Arousing Cognitive Dissonance," *Journal of Personality and Social Psychology*, 16, 199–206.

COOPER, J., M. P. ZANNA, AND P. A. TAVES (1975): "On the Necessity of Arousal for Attitude Change in the Induced Compliance Paradigm," Unpublished Manuscript, Princeton University.

COYNE, J. C., AND I. H. GOTLIB (1983): "The Role of Cognition in Depression: A Critical Appraisal," *Psychological Bulletin*, 94, 472–505.

CROYLE, R., AND J. COOPER (1983): "Dissonance Arousal: Physiological Evidence," *Journal of Personality and Social Psychology*, 45, 782–791.

DAVIS, K. E., AND E. E. JONES (1960): "Changes in Interpersonal Perception as a Means of Reducing Cognitive Dissonance," *Journal of Abnormal and Social Psychology*, 61, 402–410.

DECI, E. (1972): "The Effects of Contingent and Non-contingent Rewards and Controls on Intrinsic Motivation," *Organizational Behavior and Human Performance*, 8.

DIVISION OF CANCER PREVENTION AND CONTROL (1986): "Annual Cancer Statistics Review," Discussion paper, National Cancer Institute.

FACIONE, N. C. (1993): "Delay Versus Help Seeking for Breast Cances Symptoms: A Critical Review of the Literature on Patient and Provider Delay," *Social Science and Medicine*, 36(12), 1521–1534.

FESTINGER, L. (1957): *A Theory of Cognitive Dissonance.* Stanford University Press.

FISKE, S. T., AND S. E. TAYLOR (1991): *Social Cognition.* McGraw-Hill, 2nd edn.

FREY, D. (1986): "Recent Research on Selective Exposure to Information," in *Advances in Experimental Social Psychology*, vol. 19, pp. 41–80. Academic Press.

FREY, D., AND R. WICKLUND (1978): "A Clarification of Selective Exposure: The Impact of Choice," *Journal of Experimental Social Psychology*, 14, 132–139.

GAES, G. G., V. MELBURG, AND J. T. TEDESCHI (1986): "A Study Examining the Arousal Properties of the Forced Compliance Situation," *Journal of Experimental Social Psychology*, 22, 136–147.

GEANAKOPLOS, J., D. PEARCE, AND E. STACCHETTI (1989): "Psychological Games and Sequential Rationality," *Games and Economic Behavior*, 1(1), 60–79.

GERARD, J. B. (1967): "Choice, Difficulty, Dissonance, and the Decision Process," *Journal of Personality*, 35, 91–108.

GERVAIS, S., AND T. ODEAN (1999): "Learning To Be Overconfident," Mimeo, UC Davis.

GOETHALS, G. R., J. COOPER, AND A. NAFICY (1979): "Role of Foreseen, Foreseeable, and Unforeseeable Behavioral Consequences in the Arousal of Cognitive Dissonance," *Journal of Personality and Social Psychology*, 37, 1179–1185.

GREENWALD, A. G. (1980): "The Totalitarian Ego: Fabrication and Revision of Personal History," *American Psychologist*, 35(7), 603–618.

GREENWALD, A. G., AND S. J. BRECKLER (1985): "To Whom is The Self Presented?," in *The Self and Social Life*, ed. by B. Schlenker. New York: McGraw-Hill.

HARTER, S. (1985): "Competence as a Dimension of Self-Evaluation," in *The Development of the Self*, ed. by R. L. Leahy, chap. 2, pp. 55–121. Academic Press.

HEIDER, F. (1958): *The Psychology of Interpersonal Relations.* Wiley.

HOLMSTRÖM, B., AND P. MILGROM (1991): "Multi-Task Principal-Agent Analyzes: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics and Organization*, 7, 24–52, special issue.

JOVANOVIC, B. (1982): "Selection and the Evolution of Industry," *Econometrica*, 50(3), 649–670.

KAHNEMAN, D., P. SLOVIC, AND A. TVERSKY (1982): *Judgment Under Uncertainty: Heuristics and Biases.* Cambridge and New York: Cambridge University Press.

KREPS, D. M., AND E. L. PORTEUS (1978): "Temporal Resolution of Uncertainty and Dynamic Choice Theory," *Econometrica*, 46(1), 185–200.

KRUEGER, A., AND L. SUMMERS (1988): "Efficiency Wages and the Interindustry Wage Structure," *Econometrica*, 56, 259–293.

KUNDA, Z. (1987): "Motivated Inference: Self-Serving Generation and Evaluation of Causal Theories," *Journal of Personality and Social Psychology*, 53(4), 636–647.

LAIBSON, D. (1996): "Hyperbolic Discount Functions, Undersaving, and Savings Policy," *National Bureau of Economic Research Working Paper.*

——— (1997): "Hyperbolic Discounting and Time Preference Heterogeneity," *Harvard University Mimeo.*

LANDIER, A. (1999): "Wishful Thinking: A Model of Overconfidence and Optimal Reality Denial," Mimeo in progress.

LARRICK, R. P., AND T. L. BOLES (1995): "Avoiding Regret in Decisions with Feedback: A Negotiation Example," *Organizational Behavior and Human Decision Processes*, 63(1), 87–97.

LEA, S. E. G., AND P. WEBLEY (1997): "Pride in Economic Psychology," *Journal of Economic Psychology*, 18, 323–340.

LERMNA, C., C. HUGHES, S. LEMON, D. MAIN, C. SNYDER, C. DURHAM, S. NAROD, AND H. LYNCH (1998): "What You Don't Know Can Hurt You: Adverse Psychological Effects in Members of BRCA1-Linked and BRCA2-Linked Families Who Decline Genetic Testing," *Journal of Clinical Oncology*, 16, 1650–1654.

LORD, C. G., L. ROSS, AND M. LEPPER (1979): "Biased Assimilation and Attitude Polarization: The Effect of Prior Theories on Subsequently Considered Evidence," *Journal of Personality and Social Psychology*, 37(11), 2098–2109.

MARKUS, H., AND Z. KUNDA (1986): "Stability and Malleability of the Self-Concept," *Journal of Personality and Social Psychology*, 51(4), 858–866.

MCFARLAND, C., AND M. ROSS (1982): "Impact of Causal Attributions on Affective Reactions to Success and Failure," *Journal of Personality and Social Psychology*, 43(5), 937–946.

MEYER, H. H. (1975): "The Pay-for-Performance Dilemma," *Organizational Dynamics*, 3(3), 39–65.

MISCHEL, W., AND E. STAUB (1965): "Effects of Expectancy on Working and Waiting for Longer Rewards," *Journal of Personality and Social Psychology*, pp. 625–633.

MOR, V., S. MASTERSON-ALLEN, R. GOLDBERG, E. GUADAGNOLI, AND M. S. WOOL (1990): "Pre-diagnostic Symptom Recognition and Help Seeking Among Cancer Patients," *Journal of Community Health*, 15(4), 253–266.

MOYER, A., AND E. G. LEVINE (1998): "Clarification of the Conceptualization and Measurement of Denial in Psychosocial Oncology Research," *Annals of Behavioral Medicine*, 20(3), 149–160.

NEL, E., R. HELMREICH, AND E. ARONSON (1969): "Opinion Change in the Advocate as a Function of the Persuasibility of His Audience: A Clarification of the

Meaning of Dissonance," *Journal of Personality and Social Psychology*, 12, 117–124.

ODEAN, T. (1998): "Are Investors Reluctant to Realize Their Losses?," Mimeo, UC Davis.

O'DONOGHUE, T., AND M. RABIN (1999a): "Doing It Now Or Later," *American Economic Review*, 89(1), 103–124.

——— (1999b): "Procrastination in Preparing for Retirement," in *Behavioral Dimensions of Retirement Economics*, ed. by H. J. Aaron, pp. 125–156. Brookings Institution Press.

PALLAK, M. S., AND T. S. PITTMAN (1972): "General Motivation Effects of Dissonance Arousal," *Journal of Personality and Social Psychology*, 21, 349–358.

PEARLIN, L. I., AND C. SCHOOLER (1978): "The Structure of Coping," *Journal of Health and Social Behavior*, 19, 2–21.

PERLOFF, L. S., AND B. K. FETZER (1986): "Self-Other Judgments and the Perceived Vulnerability of Victimization," *Journal of Personality and Social Psychology*, 50, 502–510.

PETTY, R. E., AND J. T. CACIOPPO (1981): *Attitudes and Persuasion: Classic and Contemporary Approaches*. Wm. C. Brown Company Publishers.

PRYCE, J. D. (1973): *Basic Methods of Linear Functional Analysis*. Hutchison University Library, London.

RABIN, M. (1995): "Moral Preferences, Moral Constraints, and Self-Serving Biases," Working Paper.

RABIN, M., AND J. SCHRAG (1999): "First Impressions Matter: A Model of Confirmatory Bias," *Quarterly Journal of Economics*, 114(1), 37–82.

RIESS, M., P. ROSENFELD, V. MELBURG, AND J. T. TEDESCHI (1981): "Self-Serving Attributions: Biased Private Perceptions and Distorted Public Descriptions," *Journal of Personality and Social Psychology*, 41(2), 224–231.

SHAPIRO, C., AND J. STIGLITZ (1984): "Equilibrium Unemployment as a Worker Disciplinary Device," *American Economic Review*, 74, 433–444.

SIMOES, E. J., C. J. NEWSCHAFFER, N. HAGDRUP, F. ALI-ABARGHOUI, X. TAO, N. MACK, AND R. C. BROWNSON (1999): "Predictors of Compliance with Recommended Cervical Cancer Screening Schedule: A Population-Based Study," *Community Health*, 24(2), 115–130.

SPENCE, K. W., I. E. FARBER, AND H. H. MCFANN (1956): "The Relation of Anxiety (Drive) Level to Performance in Competitional Paired-Associates Learning," *Journal of Experimental Psychology*, 52, 296–305.

STAW, B. M. (1974): "Attitudinal and Behavioral Consequences of of Changing a Major Organizational Reward: A Natural Field Experiment," *Journal of Personality and Social Psychology*, 29, 742–751.

STROOCK, D. W. (1994): *A Concise Introduction to the Theory of Integration.* Birkhäuser.

SVENSON, O. (1981): "Are We Less Risky and More Skillful Than Our Fellow Drivers?," *Acta Psychologica*, 18, 473–474.

TAYLOR, S. E. (1983): "Adjustment to Threatening Events," *American Psychologist*, pp. 1161–1173.

TAYLOR, S. E., AND J. D. BROWN (1988): "Illusion and Well-Being: A Social Psychological Perspective on Mental Health," *Psychological Bulletin*, 103(2), 193–210.

TETLOCK, P. E., AND A. LEVI (1982): "Attribution Bias: On the Inconclusiveness of the Cognition-Motivation Debate," *Journal of Experimental Social Psychology*, 18, 68–88.

VISCUSI, W. K. (1994): "Cigarette Taxation and the Social Consequences of Smoking," *NBER Conference on Tax Policy and the Economy.*

WEINBERG, B. A. (1999): "A Model of Overconfidence," Mimeo, Ohio State University.

WEINSTEIN, N. D. (1980): "Unrealistic Optimism About Future Life Events," *Journal of Personality and Social Psychology*, 39, 806–820.

WELLS, L. E., AND P. D. SWEENEY (1986): "A Test of Three Models of Bias in Self-Assessment," *Social Psychology Quarterly*, 49, 1–10.

WOOL, M. S. (1986): "Extreme Denial in Breast Cancer Patients and Capacity for Object Relations," *Psychotherapy and Psychosomatics*, 46(4), 196–204.

ZÁBOJNIK, J. (1999): "A Theory of Rational Bias in Self-Assessment," Mimeo.

ZANNA, M. P., AND J. COOPER (1976): "Dissonance and the Attribution Process," in *New Directions in Attribution Research*, ed. by J. H. Harvey, W. J. Ickes, and R. F. Kidd, pp. 199–217. L. Erlbaum Associates.