University of Pennsylvania
ScholarlyCommons

Business Economics and Public Policy Papers

Wharton Faculty Research

5-2014

Imperfect Memory and Choice Under Risk

Daniel A. Gottlieb
University of Pennsylvania

Follow this and additional works at: https://repository.upenn.edu/bepp_papers

Part of the [Behavioral Economics Commons](#)

Recommended Citation

Gottlieb, D. A. (2014). Imperfect Memory and Choice Under Risk. *Games and Economic Behavior*, 85 127-158. <http://dx.doi.org/10.1016/j.geb.2013.11.013>

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/bepp_papers/118
For more information, please contact repository@pobox.upenn.edu.

Imperfect Memory and Choice Under Risk

Abstract

This paper presents a model of choice based on imperfect memory and self-deception. I assume that people have preferences over their own attributes (e.g., skill, knowledge, or competence) and can manipulate their memories. The model provides a prior-dependent theory of regret aversion and allows for prior-dependent information attitudes. It implies that behavior will converge to the one predicted by expected utility theory after a choice has been faced a sufficiently large number of times.

Keywords

memory, self deception, behavioral economics

Disciplines

Behavioral Economics

IMPERFECT MEMORY AND CHOICE UNDER RISK¹

DANIEL GOTTLIEB

Department of Economics, MIT, daniel_g@mit.edu

Abstract

This paper proposes a model of choice under risk based on imperfect memory and self-deception. The model assumes that people have preferences over their own attributes and can, to some extent, manipulate their memories. It leads to a non-expected utility representation and provides a unified explanation for several empirical regularities: non-linear probability weights, small-stakes risk aversion, regret and the competence hypothesis. It also leads to endowment and sunk cost effects. The model implies that behavior will converge to the one predicted by expected utility theory after a choice has been made a sufficiently large number of times.

1 Introduction

Choices with uncertain outcomes are an important part of a person's life. The outcomes often depend on the person's own attributes (e.g., skill, knowledge, or competence) and, therefore, affect the individual's self-views. Choices that turn out to be wrong typically lead to self-doubt, while choices that turn out to be right enhance the person's self-image. Hence, a person who cares about self-image has an incentive to manipulate recollections and beliefs. Indeed, there is sizeable psychological evidence that people value a positive self-image and manipulate their memories (see Section 2).

This paper analyzes how the concern for self-image affects an individual's behavior under risk when memory is imperfect. The model is based on two basic premises: First, individuals have preferences over their own attributes; Second, they can (to some extent) affect what they will remember. Both assumptions are largely supported by evidence from the psychology literature. Apart from these two assumptions, individuals are assumed to behave as in standard economic models. Their preferences satisfy the axioms of expected utility theory. Furthermore, individuals follow Bayes' rule and, therefore, are aware of their memory imperfection. The model ties the concept of self-deception together with several deviations from

¹I thank Muhamet Yildiz for insightful guidance and detailed suggestions, and Bengt Holmstrom and Drazen Prelec for valuable comments and suggestions. I also thank Eduardo Azevedo, Abhijit Banerjee, Roland Benabou, Moshe Cohen, Mathias Dewatripont, Peter Eso, Xavier Gabaix, Lucas Maestri, Jean Tirole, Birger Wernerfelt, and seminar participants at FGV, Fucape, MIT, and PUC-Rio for comments.

standard expected utility theory, such as ambiguity aversion, non-linear probability weights, risk aversion over lotteries with small stakes, regret aversion and the competence hypothesis. It also leads to endowment and sunk cost effects.

In its simplest version, the model consists of a two-period decision problem. In the first period, an individual observes the realization of a signal $\sigma \in \{H, L\}$, which is informative about her attributes. Then, she chooses the probability of remembering the realization of the signal by engaging in memory manipulation. In the second period, the individual applies Bayes' rule to her recollection of the signal. Because Bayes' rule implies that, on average, the individual's interpretation of her recollections are correct, self-deception does not change her (ex-ante) expected self-views. Hence, from an ex-ante point of view, memory manipulation is wasteful and, therefore, the agent would prefer not to observe the realization of the signal. Nevertheless, after observing the signal, the individual has an incentive to manipulate her memory in order to improve her self-image.

The model leads directly to preferences for avoiding information: people prefer not to acquire certain information if the expected benefit from making an informed decision is lower than the costs of self-deception. Because individuals anticipate these costs, they may prefer to make uninformed decisions if the objective value of information is sufficiently low. This result contrasts with Blackwell's celebrated theorem, which states that additional information can never be harmful. It is consistent, however, with the large psychology literature that connects self-deception and information avoidance. For example, people may avoid health exams, especially if the value of information is not high enough (e.g. the disease is not easily treatable) and if being diagnosed with the disease significantly affects the person's self-image. Individuals may also engage in "self-handicapping" strategies, such as under-preparing for an examination or getting too little sleep before physical exercise, in order to reduce the informational content of the signal. They may also display a "fear of competition" since outcomes from competitors are often informative about the person's own attributes.

When outcomes $\sigma \in \{H, L\}$ consist of monetary payments, the individual's expected utility can be represented by

$$w(q)u_H + [1 - w(q)]u_L,$$

where u_s is the decision-maker's utility in the state where s occurs and q is the probability of state $s = H$. The probability weight $w(q)$ is lower than the actual probability q when outcomes lead to memory manipulation. Hence, these preferences provide a self-deception explanation for non-expected utility and ambiguity aversion.

As in other models that admit non-expected utility representations, the decision-maker may reject gambles with small but positive expected value. The agent may also exhibit a gap between the maximum willingness to pay for a good and the minimum compensation demanded for the same good (endowment effect). However, unlike other non-expected utility models, the departure from linear weights in my model is directly related to the decision-maker's self-perceived attributes. This departure is consistent with experimental evidence suggesting that deviations from expected utility theory are associated with the lotteries' being correlated with the decision-maker's skill or knowledge (c.f., Heath and Tversky, 1991, Josephs et al., 1992, Fox and Tversky, 1995, Goodie, 2003, and Goodie and Young, 2007).² In particular, the model provides a formalization of the (informal) theory of regret aversion based on self-perception proposed by Josephs et al. (1992). According to this theory, individuals

²See Subsection 5.2 for a more detailed discussion.

with low self-image are more likely to make choices that minimize the possibility of regret. While different patterns may also be consistent with the model, it is able to predict the behavior described by Heath and Tversky (1991), according to which individuals prefer a knowledge-based lottery instead of a knowledge-independent lottery *with the same expected probability of winning* if and only if the individual believes that the probability of a positive outcome is high (competence hypothesis).³ The model also allows the decision maker to reject small gambles without imposing unrealistic degrees of risk aversion over large gambles.

Two applications illustrate the theory. Successful trading usually requires certain skills or knowledge. At the very least, the agent must form expectations about how much each good is worth. In more complex markets, future prices of the goods must also be estimated. Thus, the outcome of the trade is informative about the person's skills or knowledge. Since decision-makers avoid information correlated with skills or knowledge, they will accept a trade only if the expected benefit from the trade exceeds a certain positive threshold. Therefore, self-deception leads to an endowment effect.

The second application considers the influence of sunk decisions on behavior. In several contexts, revising one's decision usually involves admitting that a wrong decision was made and, therefore, it is often informative to the person about her own skills or knowledge. My model provides a self-deception explanation for the influence of sunk decisions on behavior that is consistent with arguments from the literature in psychology.

In a repeated setting in which the person observes a sequence of signals and engages in memory manipulation after each signal is realized, the attitude towards risk converges to the one implied by expected utility theory. This result is consistent with the arguments that people do not exhibit ambiguity aversion over events that have been observed several times and that experts are subject to much less bias than beginners (e.g. List, 2003, List and Haigh, 2005).

The structure of the paper is as follows. Section 2 briefly reviews the psychological evidence on the memory and the related literature in economics. Section 3 introduces and discusses the general framework. In Section 4, I describe the implications for information acquisition. Section 5 considers lotteries over money and provides a representation result. In Section 6, I analyze a repeated version of the model. Section 7 presents the two applications of the model. Section 8 summarizes the main results and discusses possible extensions. The appendix relaxes some assumptions from the model and presents the proofs of the propositions in the text.⁴

2 Related Literature

2.1 An Overview of the Psychology Literature

Ego-involvement, or its absence, makes a critical difference in human behavior. When a person reacts in a neutral, impersonal, routine atmosphere, his behavior is one thing. But when he is behaving personally, perhaps excitedly, seriously

³The model is also consistent with behavior that Eliaz and Spiegler (2006) have shown to be inconsistent with the Psychological Expected Utility model of Caplin and Leahy (2001).

⁴Appendix A relaxes the additive separability assumption made in Section 5; Appendix B considers naive decision-makers, who are unaware of their memory imperfection; Appendix C considers models with any finite number of possible states, and Appendix D presents the proofs.

committed to a task, he behaves quite differently. In the first condition his ego is not engaged; in the second, it is. (Gordon W. Allport, 1943, pp. 459).

Psychologists have largely documented a human tendency to deny or misrepresent reality to oneself (i.e., engage in self-deception). In general, people consider themselves to be “smart,” “knowledgeable,” and “nice.” Information conflicting with this image is usually ignored or denied. Greenwald (1980, pp. 605), for example, argued that “[o]ne of the best established recent findings in social psychology is that people perceive themselves readily as the origin of good effects and reluctantly as the origin of ill effects.” Similarly, Gollwitzer, Earle, and Stephan (1982, pp. 702), claimed that the “asymmetrical attributions after success and failure” is a “firmly established finding.”

People are also more likely to remember successes than failures (Korner, 1950). After choosing between two different options, they tend to recall the positive aspects of the chosen option and the negative aspects of the forgone option (Mather, Shafir, and Johnson, 2003). Relatedly, individuals overestimate their achievements and readily find evidence that they possess attributes which they believe to be correlated with success in personal or professional life (Kunda and Sanitioso, 1989; Quattrone and Tversky, 1984). Success is usually attributed to one’s own ability and effort, whereas failure tends to be attributed to bad luck or other external variables (Gollwitzer, Earle, and Stephan, 1982, Zuckerman, 1979).⁵ In group settings, where each individual’s contribution cannot be unequivocally determined, people tend to attribute to themselves a larger share of the group’s outcome after a success and a smaller share after a failure (Johnston, 1967).

Self-assessments and the memory are intrinsically connected. In his *Essay Concerning Human Understanding*, Locke (1690) identified the self with memory. Mill (1829, Vol. 2, pp. 174) argued that “[t]he phenomenon of Self and that of Memory are merely two sides of the same fact.” Modern cognitive psychologists define the self as the “mental representation of oneself, including all that one knows about oneself” (Kihlstrom et al., 2002). Therefore, a model of self-views should devote considerable attention to memory.

In psychology, the memory is typically viewed as imperfect and manipulable. Rapaport (1961), for example, conceived “memory not as an ability to revive accurately impressions once obtained but as the integration of impressions into the whole personality and their revival *according to the needs of the whole personality*.” Allport (1943) believed that self-deception was a mechanism of ego defense and the maintenance of self-esteem. Hilgard (1949, pp. 374) argued that “the need for self-deception arises because of a more fundamental need to maintain or to restore self-esteem. Anything belittling the self is to be avoided.” Festinger (1957) suggested that individuals have a tendency to seek consistency among their cognitions (i.e., beliefs and opinions). He labeled the discomfort felt when one is presented with evidence that conflicts with one’s beliefs and the resulting effort to distort those beliefs or opinions cognitive dissonance. In a review of the recent literature in social psychology, Sedikides, Green, and Pinter (2004, pp. 165) describe people as “striving for a positive self-definition or the avoidance of a negative self-definition (...) at the expense of accuracy and truthfulness.” According to them, “[m]emory serves the function of shielding a positive self-definition from negativity.”

⁵Van den Steen (2004) presents a model of rational agents with differing priors that generates these biases. Harbaugh (2008) provides a career concerns explanation for prospect theory.

There are several reasons why people may want to believe in things that are not true. First, there may be a hedonic value of positive self-views so that people simply like to think that they have these attributes.⁶ Second, as argued by Compte and Postlewaite (2004), a person may benefit from having overconfident beliefs in situations where emotions affect performance. Third, manipulating one’s own beliefs may facilitate the deception of others. Thus, holding an optimistic view of oneself may help convincing others of one’s own value.⁷ Fourth, there may be a motivational value of belief manipulation. As Benabou and Tirole (2002) and Weinberg (2006) argued, confidence in one’s ability may help the person set more ambitious goals and persist in adverse situations.

This paper abstracts from the exact reason why people may value a positive self-image. The model developed here is based on the two basic ideas discussed above. First, individuals have preferences over their attributes. Second, they can affect what they will remember. The paper focuses on how memory manipulations affect the person’s attitudes towards risk.

As the opening quote from Allport demonstrates, psychologists have long realized that self-deception may change a person’s behavior. Festinger (1957, pp. 3), for example, argued that “[w]hen dissonance is present, in addition to trying to reduce it, the person will actively avoid situations and information which would likely increase the dissonance.” More recently, Josephs et al. (1992, pp. 27) argued that “[r]isky decisions are potentially threatening to self-esteem because the chosen alternative will occasionally yield a less desirable outcome than would some other alternative. When a less desirable outcome does occur, it can sometimes lead people to doubt their judgement and ability, especially when the decision is an important one.”

This paper shows that incorporating self-deception in a standard model of choice can lead to a unified theory of choice under risk that is consistent with economic phenomena such as ambiguity aversion, risk aversion over lotteries with small stakes, regret, and the competence hypothesis. It also leads to endowment and sunk cost effects.

2.2 An Overview of the Literature on Imperfect Memory

The economic literature on imperfect memory can be divided in two strands. The first assumes that decision makers are naive and act as if they have not forgotten anything (Mullainathan, 2002). The other strand assumes that decision makers are sophisticated, so that they draw Bayesian inferences given that they might have forgotten things. This paper follows the latter approach and considers the case of rational decision makers subject to imperfect recall.⁸ As suggested by Piccione and Rubinstein (1997), the resulting game of imperfect recall is solved by the principle of “multiself consistency,” whereby decisions made in different stages are viewed as being made by different incarnations of the decision maker.

Models of limited memory are a special case of imperfect memory. They were originally proposed by Robbins (1956) in the mathematical statistics literature. He suggested a decision rule for choosing between two lotteries with unknown distributions that was conditional on

⁶For example, in Schelling’s (1985) theory of the mind as a consuming organ, self-views have a hedonic value.

⁷As argued by Trivers (2000, pp. 115), “[b]eing unconscious of ongoing deception may more deeply hide the deception. Conscious deceivers will often be under the stress that accompanies attempted deception.” This argument is modelled formally by Byrne and Kurland (2001) in an evolutionary game.

⁸Appendix B considers the case of naive decision makers.

a finite number of outcomes (finite memory). In a series of papers, Cover and Hellman characterized optimal solutions to some finite memory problems.⁹ More recently, economists have independently studied optimal decision making subject to limited memory. Dow (1991) considered the behavior of a consumer looking for the lowest price. Wilson (2003) studied how limited memory leads to certain biases in belief formation. Hirshleifer and Welch (2002) considered informational cascades generated by players who observe actions but not the information leading to such actions.

In a sequence of papers, Benabou and Tirole have used imperfect memory frameworks to study questions from the psychology literature. Based on the assumption that agents recalled actions but not their motivations, they have proposed theories of personal rules and internal commitments (Benabou and Tirole, 2004), prosocial behavior (Benabou and Tirole, 2006b), and identity and taboos (Benabou and Tirole, 2006c). Using a model of self-deception, Benabou and Tirole (2002, 2006a) analyzed the provision of self-motivation and the formation of collective beliefs and ideologies.

The model of memory presented here is general enough to allow for an agnostic view of the behavior of the memory system. It encompasses both Benabou and Tirole’s self-deception framework and a static version of the limited memory framework as special cases. This paper is also connected to the economic literature on cognitive dissonance (Akerlof and Dickens, 1982, Rabin, 1994). This literature assumes that agents derive utility from their beliefs and that they can, at some cost, choose their beliefs. Separately, Lowenstein (1987), Caplin and Leahy (2001 and 2004), and Köszegi (2006) have studied models with anticipatory emotions.¹⁰

3 General Framework

3.1 The Decision Problem

The model examines a decision maker (DM) who has preferences over her attributes θ . Attributes θ may be interpreted as skills, knowledge, or competence as well as a parameter of anticipatory utility. Let Θ be a non-empty subset of $\mathbb{R}$ representing the possible values of θ and let $F(\cdot)$ denote the agent’s prior distribution of θ .¹¹

The DM acts in 3 periods ($t = 0, 1, 2$). In period 0, she chooses an action a from a non-empty, compact subset of a finite dimensional Euclidean space A . For example, a can be an investment decision or a decision of whether to undertake some medical examination. The set A can also be a singleton, in which case the agent makes no choice in period 0.¹²

In period $t = 1$, an outcome σ_a , which can be either high (H) or low (L), is observed. The outcome σ_a may be a purely informative signal, entering the agent’s preferences only

⁹See Hellman and Cover (1973) for a review of the main results in this literature.

¹⁰Brunnermeier and Parker (2005) proposed a theory of “optimal expectations,” according to which individuals choose their beliefs balancing the gains from anticipating a higher future utility with the losses from suboptimal decision-making. Similarly, Hvide (2002) proposed the notion of “pragmatic beliefs,” which are the beliefs that maximize the individual’s utility. Bernheim and Thomsen (2005) showed that memory imperfections and anticipatory emotions may lead to a resolution of Newcomb’s Paradox and sustain cooperation in the Prisoners Dilemma.

¹¹ Θ can be continuous or discrete, as long as it contains at least two elements (otherwise, θ cannot be random). Note that we have not assumed that the agent has a correct prior distribution over θ . Therefore, agents are allowed to hold optimistic or pessimistic beliefs about their attributes.

¹²In some applications, the set A may also include the possibility of not observing a signal.

indirectly through her beliefs about her attributes θ . It may also affect the agent's preferences directly. For example, a medical exam consists of a purely informative signal, whereas the outcomes of an investment affect an individual not only through their informational content but also through the different monetary payments associated with them. I denote by $q_a \in (0, 1)$ the probability of observing a high outcome given action $a \in A$. A high outcome is assumed to be more favorable than a low outcome in the sense of first-order stochastic dominance:

$$F(\theta|\sigma_a = H) \leq F(\theta|\sigma_a = L) \text{ for all } \theta \in \Theta, \quad (1)$$

with strict inequality for some value of θ , and for all $a \in A$.

Following Rabin (1994), Benabou and Tirole (2002, 2006a), and Benabou (2008), I assume that the individual can, at a cost, influence her recollections. The DM remembers the outcome $s \in \{H, L\}$ with probability

$$\eta_s + m_s,$$

where the parameter $\eta_s \in [0, 1]$ is the agent's "natural" rate of remembering outcome s . This rate determines the probability that the DM recollects the outcome if she does not employ any manipulation effort. However, the DM is also able to depart from the natural rate of forgetting the outcome by exerting effort $m_s \in [-\eta_s, 1 - \eta_s]$ in period $t = 1$. Engaging in memory manipulation m_s leads to a cost of $\psi_s(m_s) \geq 0$, $s \in \{H, L\}$. The agent's recollection of the outcome σ_a is denoted by $\hat{\sigma}_a \in \{H, L, \emptyset\}$, where we write $\hat{\sigma}_a = \emptyset$ if the outcome has been forgotten.

In period $t = 2$, the DM takes an action b in a non-empty, compact subset of a finite dimensional Euclidean space B . For example, b can be a decision of whether to continue with some previous investment or whether to undertake some medical treatment. B can also be a singleton, in which case the DM does not act after observing the outcome. Figure 1 presents the informational structure.

Preferences satisfy the standard axioms of expected utility theory. Therefore, there exists utility function $u : \Theta \times A \times B \times \{H, L\} \rightarrow \mathbb{R}$ representing the DM's preferences. Furthermore, $u(\theta, a, b, \sigma)$ is strictly increasing in θ for all $(a, b, \sigma) \in A \times B \times \{H, L\}$.

When $u(\theta, a, b, H) = u(\theta, a, b, L)$ for all $(\theta, a, b) \in \Theta \times A \times B$, we refer to outcomes as signals since they do not affect the agent's utility directly. In this case, we say that the model has *purely informative signals*. When signals are purely informative and A and B are singletons, we say that signals have a *purely hedonic value*. In models where signals have a purely hedonic value, the DM does not need to take any decision and the only reason for memory manipulation is the improvement of the individual's self-views.

We refer to the case where $u(\theta, a, b, H) > u(\theta, a, b, L)$ for all $(\theta, a, b) \in \Theta \times A \times B$ as a model of *monetary outcomes*. In this case, outcomes are interpreted as monetary payments and a high outcome raises the agent's utility both directly and through beliefs about θ .¹³

The cost of memory manipulation ψ_s can be related to psychic costs (stress from repression of negative information or effort to focus on positive information), time (searching for reassuring information or excuses, lingering over positive feedback), or real resources (avoiding certain cues and interactions or eliminating evidence). They can also be interpreted as

¹³ Although the case described above, where the outcome with a higher monetary payment provides more favorable news about the DM's attributes, is the most intuitive, this is not necessary for our results. Alternatively, one could assume that the outcome with a higher monetary payment is bad news about the DM's attributes.

Figure 1: Informational Structure

the shadow costs of memory in a limited information framework. Remembering an outcome with probability above its natural rate η_s requires an individual to focus on it and on information correlated with it. In turn, this restricts the amount of attention available to other information (which has shadow cost ψ_s). Similarly, forgetting an outcome with probability above the natural rate $1 - \eta_s$ requires an individual to focus on confronting evidence which again restricts the amount of attention available to other potentially useful information.¹⁴

Assumption 1 The cost of memory manipulation $\psi_s(m_s)$ is strictly decreasing in $m_s < 0$, strictly increasing in $m_s > 0$, convex, twice-continuously differentiable, and such that $\psi_s(0) = 0$, $s \in \{H, L\}$.

Figure 2 depicts the costs of memory manipulation implied by Assumption 1. I further assume that the agent forgets a high outcome with some positive probability if she does not exert any effort:¹⁵

Assumption 2 $\eta_H < 1$.

The model can also be seen as a conflict between a “hot” or “impulsive” self and a “cold” self. The hot self (self 1) wants to minimize current losses from negative information and maximize the current gains from positive information.¹⁶ The cold self (self 2) wants to circumvent the manipulations made by the hot self in order to make a correct inference. The hot self exerts efforts m_L and m_H in order to manipulate the beliefs of the cold self. Then,

¹⁴For example, Steele’s (1988) self-affirmation theory argues that people cope with negative outcomes in one domain by focusing in other, unrelated domains.

¹⁵If $\eta_H = 1$, then the model becomes trivial. Since the agent always recalls high outcomes, she will perfectly infer that $\sigma = L$ was observed if she recollects $\hat{\sigma} = \emptyset$. Therefore, she will never engage in memory manipulation.

¹⁶This interpretation assumes that the hot self is rational in the sense of taking into account the benefits and costs of memory manipulation. Several papers in social psychology have documented that individuals tend to be more realistic and impartial when making important decisions (c.f., Taylor and Gollwitzer, 1995, and references therein). Therefore, self-deception seems to decrease when the cost of a mistake increases. Prelec (2008) presented experimental evidence where self-deception responds positively to its expected benefits.

Similarly to this interpretation, Bodner and Prelec (2002) present a signaling model between an agent’s privately informed gut and the agent’s uninformed mind.

Figure 2: Cost of Memory

the cold self applies Bayes' rule in order to filter these manipulations and make a correct decision b .¹⁷

As the following examples show, the general framework encompasses other models of imperfect memory.

Example 1 (The Forgetfulness Model of Benabou and Tirole, 2002) Take $\eta_L = 1$, $\eta_H = 0$ and $\psi_H(m_H) = +\infty$ for all $m_H > 0$ so that high outcomes are always forgotten (i.e., $\eta_H + m_H = 0$). Figure 3 presents the informational structure in this case. This is the memory framework from Benabou and Tirole (2002). It can be interpreted as a model of bad news or no news. If the agent receives bad news, she can exert an effort $m_L \in [-1, 0]$ in order to forget them.

If the state $\emptyset$ is reinterpreted as the recollection of a high outcome, then the model from Example 1 becomes one where the agent is able to convince herself that a low outcome was a high outcome.¹⁸ Hence, memory manipulation would allow the DM to believe that she observed an outcome $\sigma = H$. This reinterpretation is compatible with neurological evidence from Prelec (2008), who showed that subjects experience heavy brain activity only when they try to convince themselves that a bad outcome was actually a good one. In the other states (both when they acknowledge a mistake or when they believe to have been correct), no such activity is detected. Hence, Example 1 can be interpreted as the agent incurring psychological costs when she tries to convince herself that a bad outcome was actually a good one.

Example 2 (The Limited Memory Model) Take $\eta_L = \eta_H = 0$ so that the DM forgets any outcome if she does not employ memory efforts. Then, the framework becomes a model of limited memory. In this model, the DM must allocate a limited amount of memory in

¹⁷The model can be interpreted as a formalization of the neurophysiological argument put forth by Trivers (2000). According to this interpretation, self 1 would be the person's unconscious process of information manipulation. In the context of intertemporal choice, several papers have proposed dual self models (c.f. Thaler and Shefrin, 1981, Fudenberg and Levine, 2006, and Brocas and Carrillo, 2008).

¹⁸In this model, the agent would never choose to believe that a high outcome was actually low.

Figure 3: Forgetfulness Model

order to store information. By spending a memory cost $\psi_s(m_s)$, she remembers an outcome $s \in \{H, L\}$ with probability m_s . A higher effort m_s can be interpreted as having greater memory resources used to store the information.¹⁹

The following examples present applications of the general framework to specific environments:

Entrepreneurship Example An employed individual is considering quitting her job and starting a new company. Building a successful company requires certain entrepreneurial skills which are unknown to the individual. Therefore, a success provides favorable news about the individual's skills. If she decides not to quit her job, the individual obtains a wage $w \in \mathbb{R}_+$ and does not learn any information about her skills.

In this paper, this situation is modeled as follows. Let the individual's career choice be denoted by $a = E$ if she becomes an entrepreneur and by $a = W$ if she remains a worker and let θ denote the individual's entrepreneurial skills. The outcome from starting a company is denoted by σ , which is equal to H in the case of success and L in the case of failure. After the outcome σ is observed, the entrepreneur may engage in memory manipulation. In this model, there is no ex-post choice (B is a singleton). The agent's decision tree is presented in Figure 5.

Appendix C considers a more general model. In that model, an outcome is a vector $\sigma = (s, r)$ consisting of a binary variable reflecting whether or not the company was successful, $s \in \{S, F\}$, and an external variable $r \in \mathbb{R}$ which affects the outcomes but is independent of the agent's attributes (e.g., general market conditions, economy-wide shocks). The entrepreneur always remembers whether the company succeeded or failed but may forget the prevailing external conditions r .

¹⁹Dow (1991) considers a consumer who searches sequentially for the lowest price, but who only remembers each price as belonging to a finite number of categories. Wilson (2003) considers a decision-maker who must act after a large number of periods but whose memory is restricted to a finite number of states.

Figure 4: Limited Memory Model

Succeeding under adverse conditions provides good news about the individual's skills. Similarly, failing under favorable conditions is bad news about her skills. In this model, that the agent will manipulate her memory in order to forget positive external shocks and remember negative shocks. This result is consistent with the psychological literature described in Section 2, which shows that success is usually attributed to one's own attributes whereas failure tends to be attributed to bad luck or other external variables.

Section 5 will show that self-deception will prevent some individuals from becoming entrepreneurs even when the expected monetary payoffs from starting a new company are higher than the payoff from remaining on the previous job.

Used Car Example An individual is considering whether to purchase a used car or to use public transportation. A used car may be defective. Moreover, detecting whether the car is defective requires certain skills. Therefore, purchasing a defective car conveys unfavorable information about the buyer's skills and requires the car to be fixed. If she decides to use public transportation, no information is learned.

This situation is modeled as follows. Let $a = C$ denote the choice of purchasing a used car and let $a = PT$ denote the choice of using public transportation. Denote by $\sigma = H$ the case where the car is non-defective and $\sigma = L$ the case where it is defective. After the consumer learns that the car was defective, she may manipulate her memory in order to forget that it needed to be fixed. Similarly, if the car was non-defective, she may exert some effort to remember that the car did not need to be fixed. Assuming the memory system from the forgetfulness model of Example 1, we obtain the decision tree depicted in Figure 6. Section 5 will show that if the expected monetary benefit from buying the used car is positive but lower than the expected self-deception costs, the individual will prefer not to purchase it.

3.2 Modeling as a Multiself Game

This paper follows Piccione and Rubinstein (1997) in modeling a decision problem with imperfect memory as a game between different selves. The decision maker is treated as a collection of selves, each of them unable to control the behavior of future selves. As will be

Figure 5: Entrepreneurship Example

described in Subsection 3.3, the decision made by an agent with imperfect recall corresponds to the perfect Bayesian equilibrium (PBE) of this game between selves.²⁰

The extensive form of the multiself game is presented in Figure 7. There are two players: self 1 and self 2. Both selves have the same utility functions but different information sets. In period 0, self 1 chooses an action $a \in A$. Then, nature plays a high outcome with probability q_a and a low outcome with probability $1 - q_a$. In period 1, conditional on the outcome $s \in \{H, L\}$, self 1 decides the amount of memory manipulation m_s . Then, given the outcome s and the manipulation effort m_s , nature plays $\hat{\sigma} = s$ with probability $\eta_s + m_s$ and $\hat{\sigma} = \emptyset$ with probability $1 - \eta_s - m_s$. In period 2, self 2 observes the recollection $\hat{\sigma}$ and takes an action $b \in B$. Then, both selves get payoff $E[u(\theta, a, b, \sigma) | \hat{\sigma}] - \psi_s(m_s)$.

Because the DM has preferences over θ , she has an interim incentive to manipulate her beliefs by exerting effort m_s . However, the set of possible beliefs that an agent can hold is restricted by the assumption that recollections are interpreted according to Bayes' rule. Thus, the agent makes correct inferences about her attributes θ given her recollections $\hat{\sigma}$.

Equivalently, we can conceptualize an “inferential self” who tries to make a correct inference about the agent’s attributes given the recollections. This inferential self chooses the agent’s expected utility so as to minimize a quadratic loss function:

$$u_{\hat{\sigma}}(a, b, \sigma) = \arg \min_{\hat{u} \in \mathbb{R}} \int_{-\infty}^{\infty} [\hat{u} - u(\theta, a, b, \sigma)]^2 dF(\theta | \hat{\sigma}).$$

The solution to this program is $u_{\hat{\sigma}}(a, b, \sigma) = \int u(\theta, a, b, \sigma) dF(\theta | \hat{\sigma})$, which is the Bayes estimator of $u(\theta, a, b, \sigma)$ given the recollection $\hat{\sigma}$. Thus, by minimizing a quadratic loss

²⁰For the games considered here, the set of sequential equilibria coincides with the set of PBE.

Figure 6: Used Car Example

function, the inferential self constrains the decision-maker to be a Bayesian given her memory imperfection.

Remark 1 Denote the expected value of θ conditional on the observed outcome σ_a by θ_{σ_a} and the expected attributes conditional on the recollection $\hat{\sigma}_a$ by $\hat{\theta}_{\hat{\sigma}_a}$. $\hat{\theta}_{\hat{\sigma}_a}$ is “less variable” than θ_{σ_a} in the sense of second-order stochastic dominance.²¹ Therefore, because θ_{σ_a} is the Bayes estimate of θ given the outcome σ_a , forgetfulness implies that the decision-maker updates observed outcomes σ_a less than implied by Bayes’ rule. This result is consistent with experimental evidence from Falk, Huffman, and Sunde (2006).

3.3 Solution Concept

As described in the previous subsection, the decision made by an agent with imperfect recall is modeled as the perfect Bayesian equilibrium (PBE) of the multiself game. Let $\mu(\cdot|\hat{\sigma})$ denote the DM’s posterior beliefs about θ given $\hat{\sigma}$ and let $E_\mu[\cdot|\hat{\sigma}]$ denote the expectation operator with respect to $\mu(\cdot|\hat{\sigma})$. Given a profile of memory manipulation manipulation (m_L, m_H) , let $E_{\hat{\sigma}_a}[\cdot|m_L, m_H]$ denote the expectation with respect to the distribution of $\hat{\sigma}_a$.

Definition 1 A PBE of the game is a strategy profile $(a^*, b^*, m_H^*(a), m_L^*(a))$ and posterior beliefs $\mu(\cdot|\hat{\sigma}_a)$ such that:

$$1. a^* \in \arg \max_{a \in A} \left\{ \begin{array}{l} E_{\hat{\sigma}_a} [E_\mu [u(a, b_a^*(\hat{\sigma}_a), \theta, \sigma_a) | \hat{\sigma}_a] | m_L^*(a), m_H^*(a)] \\ - q\psi_H(m_H^*(a)) - (1 - q)\psi_L(m_L^*(a)) \end{array} \right\};$$

²¹See Appendix D for the proof.

Figure 7: Extensive Form

2. $m_s^*(a) \in \arg \max_{m_s} \left\{ \begin{aligned} &(\eta_s + m_s) E_\mu [u(a, b_a^*(\hat{\sigma}_a), \theta, s) | \hat{\sigma}_a = s] \\ &+ (1 - \eta_s - m_s) E_\mu [u(a, b_a^*(\hat{\sigma}_a), \theta, s) | \hat{\sigma}_a = \emptyset] - \psi_s(m_s) \end{aligned} \right\},$
 $s \in \{H, L\};$
3. $b_a^*(\hat{\sigma}) \in \arg \max_{b \in B} \{E_\mu [u(a, b, \theta, \sigma_a) | \hat{\sigma}_a = \hat{\sigma}]\};$
4. $\mu(\theta | \hat{\sigma}_a = \hat{\sigma})$ is obtained by Bayes' rule if $\Pr(\hat{\sigma}_a = \hat{\sigma} | m_L^*(a^*), m_H^*(a^*)) > 0, \forall \hat{\sigma} \in \{L, H, \emptyset\}.$

Conditions 1 – 3 are the standard sequential rationality conditions. Condition 1 states that self 1 chooses an ex-ante action a that maximizes the agent's expected utility in period 0 given the behavior of self 2. Condition 2 states that, conditional on each outcome $s \in \{H, L\}$, self 1 chooses the amount of manipulation that maximizes her expected payoff. Condition 3 states that self 2 takes an action that maximizes her utility given the beliefs she holds about the manipulation employed by self 1.

Condition 4 is the standard consistency condition, requiring that beliefs of self 2 satisfy Bayes' rule given the strategy of self 1. For every recollection $\hat{\sigma}$ that is reached with positive probability, it implies that $\mu(\theta|\hat{\sigma}_a) = F(\theta|\hat{\sigma}_a)$. Because of Bayesian updating, Condition 3 becomes

$$b_a^*(\hat{\sigma}) \in \arg \max_b \int u(a, b, \theta, \sigma) dF(\theta|\hat{\sigma}_a = \hat{\sigma}),$$

for any recollection $\hat{\sigma}$ that is reached with positive probability. The following proposition establishes the existence of a PBE:

Proposition 1 (Existence) *There exists a PBE.*

Define the expected utilities given $\sigma_a = H$ and $\sigma_a = L$ by

$$\begin{aligned} u_H(a, b, \sigma_a) &\equiv \int u(a, b, \theta, \sigma_a) dF(\theta|\sigma_a = H), \text{ and} \\ u_L(a, b, \sigma_a) &\equiv \int u(a, b, \theta, \sigma_a) dF(\theta|\sigma_a = L). \end{aligned} \quad (2)$$

Given the recollection of a high signal, $\hat{\sigma}_a = H$, self 2 infers that a high signal was observed in period 1. Hence, Bayesian updating implies that the expected utility of self 1 conditional on $\hat{\sigma}_a = H$ is $u_H(a, b_a(H), H)$. Similarly, the expected utility of self 1 conditional on $\hat{\sigma}_a = L$ is $u_L(a, b(L), L)$.

Let $m_L^*(a)$ and $m_H^*(a)$ denote the amount of memory manipulation that self 2 believes was employed in period 1. Note that the PBE concept implies that $m_L^*(a)$ and $m_H^*(a)$ are taken as given by self 1 when choosing the amount of memory manipulation to exert. If the DM forgets which signal was observed in period 1 (i.e., she recollects $\hat{\sigma}_a = \emptyset$), then there is a probability $(1 - q_a)(1 - \eta_L - m_L^*(a))$ that $\sigma_a = L$ was observed and a probability $q_a(1 - \eta_H - m_H^*(a))$ that $\sigma_a = H$ was observed. Thus, the expected utility given $\hat{\sigma}_a = \emptyset$ is

$$\begin{aligned} u_{\emptyset}(a, b_a(\emptyset), \sigma_a) &\equiv \alpha(m_L^*, m_H^*) u_H(a, b_a(\emptyset), \sigma_a) \\ &\quad + [1 - \alpha(m_L^*, m_H^*)] u_L(a, b_a(\emptyset), \sigma_a), \end{aligned} \quad (3)$$

where $\alpha(m_L, m_H) \equiv \frac{q_a(1 - \eta_H - m_H)}{q_a(1 - \eta_H - m_H) + (1 - q_a)(1 - \eta_L - m_L)}$ is the conditional probability of $\sigma_a = H$ implied by Bayes' rule.

Conditions 2 and 3 from Definition 1 state that, after observing signal $\sigma_a = s \in \{H, L\}$, self 1 chooses m_s to maximize

$$(\eta_s + m_s) u_s(a, b_a(s), s) + (1 - \eta_s - m_s) u_{\emptyset}(a, b_a(\emptyset), s) - \psi_s(m_s).$$

Using equation (3), the expected utility after a low signal can be written as

$$\begin{aligned} &u_L(a, b_a(\emptyset), L) + (\eta_L + m_L) [u_L(a, b_a(L), L) - u_L(a, b_a(\emptyset), L)] \\ &+ (1 - \eta_L - m_L) \alpha(m_L^*(a), m_H^*(a)) [u_H(a, b_a(\emptyset), L) - u_L(a, b_a(\emptyset), L)] - \psi_L(m_L) \end{aligned} \quad (4)$$

Note that self 1 takes three factors into account when choosing the amount of effort to forget bad news. First, forgetting a low signal leads to a higher utility through a more favorable inference about θ since $u_H(a, b_a(\emptyset), L) > u_L(a, b_a(\emptyset), L)$ (self-deception factor). Second, it

leads to a sub-optimal choice of b since $u_L(a, b_a(L), L) \geq u_L(a, b_a(\emptyset), L)$ (decision-making factor). Third, self-deception leads to a memory cost of $\psi_L(m_L)$ (memory cost factor).

Analogously, conditional on a high signal, self 1 chooses m_H to maximize:

$$(\eta_H + m_H) \left\{ \begin{array}{l} [1 - \alpha(m_L^*(a), m_H^*(a))] \\ + u_H(a, b_a(H), H) - u_H(a, b_a(\emptyset), H) \end{array} \right\} + u_{\emptyset}(a, b_a(\emptyset), H) - \psi_H(m_H). \quad (5)$$

This equation displays the three factors that determine the amount of effort to remember good news. First, remembering a high signal leads to a higher utility through a more favorable inference about θ since $u_H(a, b_a(\emptyset), H) > u_L(a, b_a(\emptyset), H)$. It also leads to better decision-making since $u_H(a, b_a(H), H) > u_H(a, b_a(\emptyset), H)$. However, it leads to a memory cost of $\psi_H(m_H)$.

The improvement in decision-making leads the DM to engage in an effort to remember a high signal. The effect from self-image also leads the DM to exert an effort to remember the high signal. Because small amounts of memory manipulation have second-order costs, the DM always remembers a high signal with probability above her natural rate η_H :

Proposition 2 (Remembering Good News) *Suppose that ψ_s is strictly convex, $s \in \{H, L\}$. Then, in any PBE, $m_H^*(a) > 0 \forall a \in A$.*

The DM's ex-ante expected utility (in period 0) is

$$E_{\hat{\sigma}_a} [E_{\mu} [u(\theta, a, b_{a^*}(\hat{\sigma}), \sigma) | \hat{\sigma}] - q\psi_H(m_H^*(a^*)) - (1 - q)\psi_L(m_L^*(a^*))]. \quad (6)$$

As in other decision problems with imperfect recall, the timing of decisions has important implications for the solution. If the agent could commit to a strategy at an ex-ante stage, she would generally choose a different amount of memory manipulation.

Consider, for example, the model of purely hedonic signals. In this case, equations (4) and (5) imply that the DM faces a trade-off between self-deception and memory costs. Manipulating one's memory into forgetting a low signal directly increases the individual's expected payoff by raising the probability that the signal is forgotten. Similarly, exerting effort to remember a high signal directly raises her expected payoff by decreasing the probability that the signal is forgotten. However, these manipulations also decrease the DM's expected payoff indirectly by reducing the relative probability of a high signal when the signal is forgotten. Bayesian updating implies that the indirect effects exactly cancel the direct effects out. Because the DM is not fooled on average, she adjusts the expected attributes given $\hat{\sigma} = \emptyset$ to take into account the relative frequency that each signal is forgotten. Therefore, from an ex-ante perspective, memory manipulation only leads to memory costs and the DM would prefer not to engage in memory manipulation at all ($m_H = m_L = 0$). However, the multiself approach implies that self 1 does not take into account the indirect effects from memory manipulation and, therefore, chooses to engage in memory manipulation.²² Hence, unlike in decision problems with perfect recall where ex-ante optimal strategies are always time-consistent, the ex-ante optimal strategy is time-inconsistent.²³

²²Note that the DM would never choose to undo the memory manipulation in period $t = 2$ and find out the true outcome σ if she had a chance to do so.

²³See Piccione and Rubinstein (1997) for a discussion of decision problems with imperfect recall. In the present model, because all nodes are reached with positive probability, the two equilibrium concepts proposed there (multiself consistency and modified multiself consistency) coincide.

In cases where outcomes affect ex-post actions b (i.e., information has positive value), it is ex-ante optimal to choose some positive amount of memory manipulation.²⁴ In these cases, the optimal strategy from an ex-ante perspective would always have a probability to remember (weakly) above the natural rate η_s , $s \in \{L, H\}$.

Recall that self 1 takes three factors into account when choosing the amount of memory manipulation: (i) self-deception, (ii) decision-making, and (iii) memory costs. As discussed previously, Bayesian updating implies that the self-deception effect vanishes from the DM's ex-ante utility. Since only factors (ii) and (iii) would be taken into account, the DM would choose to remember good news and to forget bad news less frequently if she could commit to a strategy in period 0. Let the ex-ante expected utility be denoted by

$$\begin{aligned} \mathcal{U}\left(m_H, m_L, a, \{b(\hat{\sigma})\}_{\hat{\sigma} \in \{H, L, \emptyset\}}\right) &= E_{\hat{\sigma}_a} [E_\mu [u(\theta, a, b(\hat{\sigma}), \sigma) | \hat{\sigma}]] \\ &\quad - q\psi_H(m_H) - (1 - q)\psi_L(m_L). \end{aligned}$$

Proposition 3 establishes this claim formally:

Proposition 3 (Excessive Manipulation) *Let $\left(\tilde{m}_H(a), \tilde{m}_L(a), \{\tilde{b}_a(\hat{\sigma})\}_{\hat{\sigma} \in \{H, L, \emptyset\}}\right)$ be a maximizer of $\mathcal{U}$ given action a and suppose $\mathcal{U}$ is a concave function of m_H and m_L .²⁵ Then, in any PBE with manipulations $m_H^*(a)$ and $m_L^*(a)$,*

$$m_H^*(a) \geq \tilde{m}_H(a) \quad \text{and} \quad m_L^*(a) \leq \tilde{m}_L(a)$$

for all $a \in A$, with at least one of the inequalities being strict.

3.4 Equilibrium when Information has Purely Hedonic Value

In order to illustrate the impact of self-deception on choice, this subsection considers the simple case where signals are purely informative and the DM does not take any action (i.e., information has purely hedonic value). In this case, the only reason for memory manipulation is the improvement of the DM's self-views. Since remembering a low signal decreases self 1's expected utility, she would never choose to manipulate her memory in order to remember a low signal (i.e., $m_L^* \leq 0$). Analogously, she would never manipulate her memory so as to forget a high signal (i.e., $m_H^* \geq 0$).

Since, in the purely hedonic case considered in this subsection, A and B are singletons and the outcome of the signal does not enter the agent's utility directly, I omit the terms a , b , and σ_a from the DM's von Neumann-Morgenstern utility function. Let $\Delta u \equiv u_H - u_L$ denote the payoff gain by observing a high signal instead of a low signal.

Proposition 4 (Forgetting Bad News) *Suppose that ψ_s is strictly convex, $s \in \{H, L\}$. Then in any PBE, $m_H^* > 0 \geq m_L^*$. Furthermore,*

$$\begin{aligned} \Delta u &\geq \psi'_H(1 - \eta_H) \implies m_H^* = 1 - \eta_H \text{ and } m_L^* = 0; \\ \Delta u &< \psi'_H(1 - \eta_H) \implies 0 < m_H^* < 1 - \eta_H \text{ and } m_L^* < 0. \end{aligned}$$

²⁴More precisely, let $b(\hat{\sigma})|_{m_L, m_H}$ denote the action that maximizes the DM's utility given recollection $\hat{\sigma}$ and conditional on manipulation efforts m_L and m_H . Then, $b(H)|_{m_L=m_H=0} \neq b(\emptyset)|_{m_L=m_H=0}$ implies that the manipulation effort m_H that maximizes the ex-ante expected utility is strictly positive. Analogously, if $b(L)|_{m_L=m_H=0} \neq b(\emptyset)|_{m_L=m_H=0}$ then m_L that maximizes the ex-ante expected utility is strictly positive.

²⁵It is straightforward to show that $\mathcal{U}$ is always a concave function of m_H and m_L when B is a singleton.

If the marginal cost of remembering good news is lower than its marginal benefit for all $m_H \in [-\eta_H, 1 - \eta_H)$, i.e. $\psi'_H(1 - \eta_H) \leq \Delta u$, then the DM always remembers high signals. In this case, there is no point in trying to forget a low signal since the agent perfectly infers that a low signal was observed when she recollects $\hat{\sigma} = \emptyset$.

If the marginal cost of remembering good news is higher than its marginal benefit for some $m_H \in [-\eta_H, 1 - \eta_H)$, then the DM forgets high signals with positive probability. In this case, because the cost of a small amount of memory manipulation is of second-order, bad news are remembered with probability below the natural rate η_L , i.e., $m_L^* < 0$.

Next, I characterize the PBE in the forgetfulness model of Benabou and Tirole (Example 1) and in the limited memory model (Example 2) when signals have purely hedonic value.

3.4.1 The forgetfulness model of Benabou and Tirole (2002)

Consider the forgetfulness model of Example 1 and suppose that ψ_L is strictly convex. Given a low signal, self 1 solves

$$\max_{m_L \in [-1, 0]} (1 + m_L) u_L - m_L \{ \alpha(m_L^*, 0) u_H + [1 - \alpha(m_L^*, 0)] u_L \} - \psi_L(m_L). \quad (7)$$

Applying Kuhn-Tucker's theorem and substituting the equilibrium condition $m_L = m_L^*$, we obtain

$$\frac{q\Delta u}{q - (1 - q)m_L^*} = -\psi'_L(m_L^*), \quad (8)$$

in any interior equilibrium.

Let m_L^* be implicitly defined by equation (8). From the implicit function theorem, such $m_L^* \in \mathbb{R}$ exists and is unique. The following proposition characterizes the PBE and presents some comparative statics results:

Proposition 5 (Characterization) *In the forgetfulness model when signals have a purely hedonic value, there exists an essentially unique PBE.²⁶ The equilibrium manipulation effort is*

$$m_L^* = \begin{cases} \psi_L'^{-1} \left(-\frac{q\Delta u}{q - (1 - q)m_L^*} \right) & \text{if } \Delta u < -\frac{\psi'_L(-1)}{q}, \text{ and} \\ -1 & \text{if } \Delta u \geq -\frac{\psi'_L(-1)}{q}. \end{cases}$$

Furthermore, the absolute value of belief manipulation $|m_L^*|$ is:

1. increasing in the benefit of manipulation Δu (for u_L fixed),
2. decreasing in the marginal cost of manipulation, and
3. increasing in q , the probability of not observing a signal.

²⁶The PBE is essentially unique in the sense that all PBE feature the same choices of actions a and b , the same manipulation efforts m_L and m_H , and the same beliefs for all recollections that are reached with positive probability. Equilibria may diverge only with respect to beliefs at recollections that are not reached with positive probability.

The comparative statics above follows from simple cost-benefit comparisons. When the marginal benefit of self-deception is higher or the marginal cost is lower, the agent chooses to engage in more self-deception. This result is consistent with the experimental evidence presented by Prelec (2008), which suggests that self-deception is increasing in the benefits of manipulation.

Also, recall that in this model, no news is good news. Therefore, when the probability of not observing a signal q is higher, it becomes more credible that the individual has not manipulated her beliefs into forgetting a low signal. Hence, an increase in q increases the marginal benefit of self-deception, and this in turn leads to an increase in the amount of memory manipulation $|m_L^*|$.

3.4.2 The limited memory model

Consider the limited memory model of Example 2. Given a high signal, self 1 solves

$$\max_{m_H \in [-1, 1]} m_H u_H + (1 - m_H) \{ \alpha(0, m_H^*) u_H + [1 - \alpha(0, m_H^*)] u_L \} - \psi_H(m_H).$$

Proceeding as in Proposition 5, it follows that the set of PBE efforts are characterized by

$$\frac{(1 - q) \Delta u}{1 - q + q(1 - m_H^*)} = \psi'_H(m_H^*), \quad (9)$$

if $\Delta u \leq \psi'_H(1) \left[1 + \frac{q}{1-q} (1 - m_H^*) \right]$, and

$$m_H^* = 1 \text{ if } \Delta u \geq \psi'_H(1).$$

Since both sides of equation (9) are increasing in m_H^* , there may be multiple interior equilibria. It may also simultaneously feature interior equilibria and a corner equilibrium.²⁷ A person that believes she often forgets good signals is not hurt much by not recalling a good signal. Therefore, she will not manipulate her memory enough and, in equilibrium, she will often forget good signals. On the other hand, a person that usually remembers good signals is severely hurt by recollecting $\hat{\sigma} = \emptyset$. Therefore, she will have more incentive to remember good signals. As I show in the next section, these equilibria are welfare ranked (from an ex-ante perspective): The equilibrium with the lowest amount of memory manipulation is preferred. The individual may be caught in a self-trap where she exerts more manipulation effort because self 1 believes that she will have engaged in more memory manipulation.²⁸

4 Purely Informative Signals and Information Acquisition

Suppose the decision-maker can choose whether or not to observe an informative signal. When would she prefer to observe it? This section is concerned with the implications of

²⁷For example, if $\psi'_H(1) \leq \Delta u \leq \psi'_r(1) \left[1 + \frac{q}{1-q} (1 - m_H^*) \right]$, there exist both an equilibrium with $m_H^* = 1$ and an interior equilibrium with m_H^* implicitly defined by equation (9).

²⁸The existence of multiple equilibria is interesting since there seems to be a large heterogeneity in the amount of self-deception across different people (c.f., Prelec, 2008). However, since the results presented here hold in all PBE, they would also be obtained if one applied a selection criterion.

memory manipulation for the acquisition of information. I show that the DM will only observe a signal if the benefit of making an informed decision exceeds the cost of memory manipulation. Subsection 4.1 discusses a theory of regret aversion based on self-deception. Then, Subsection 4.2 shows that the model is consistent with intuitive behavior that Eliaz and Spiegel (2006) have shown to be incompatible with Caplin and Leahy's (2001) Psychological Expected Utility model.

The standard theory of information acquisition under expected utility states that it is optimal to observe a signal when the value of information (defined as the expected payoff gain by observing the signal) is greater than the cost of information. Similarly, I will show that the DM prefers to observe a signal if the (objective) value of information is greater than the expected cost of self-deception. In particular, when information has purely hedonic value, the DM always prefers not to observe any signal.

The *objective value of information* is defined as the expected payoff from observing the signal:

$$V \equiv E_{\hat{\sigma}_a} [E_{\mu} [u(a, b_a(\hat{\sigma}_a), \theta) | \hat{\sigma}]] - \max_{a \in A, b \in B} \int u(a, b, \theta) dF(\theta) \geq 0, \quad (10)$$

where $\max_{a \in A, b \in B} \int u(a, b, \theta) dF(\theta)$ is the expected payoff if the DM could not observe $\hat{\sigma}_a$. Thus, equation (6) implies that the ex-ante expected utility from observing the signal $U(\Sigma)$ is equal to

$$\max_{a \in A, b \in B} \int u(a, b, \theta) dF(\theta) + V - q\psi_H(m_H^*(a^*)) - (1 - q)\psi_L(m_L^*(a^*)). \quad (11)$$

It follows that the DM would prefer to observe the signal if the objective value of information V is greater than the expected cost of memory manipulation $q\psi_H(m_H^*(a^*)) + (1 - q)\psi_L(m_L^*(a^*))$.

Proposition 6 (Information Acquisition) *Fix a PBE. Let $U(\Sigma)$ denote the expected utility of observing the signal in this PBE and let $E[u]$ denote the expected utility of not observing the signal. Then, $U(\Sigma) - E[u] = V - q\psi_H(m_H^*(a^*)) - (1 - q)\psi_L(m_L^*(a^*)) < V$.*

When information has a purely hedonic value, the objective value of information is $V = 0$. In equilibrium, when the a signal is forgotten ($\hat{\sigma} = \emptyset$), the DM knows that there is a probability $\alpha(m_L^*, m_H^*)$ that there was a high signal and $1 - \alpha(m_L^*, m_H^*)$ that there was a low signal. Bayesian updating implies that on average, the only effects of engaging in self-deception are the manipulation costs $\psi_L(m_L^*)$ and $\psi_H(m_H^*)$. Of course, there is still an interim incentive to manipulate beliefs after she observes the signal. *The inability to commit not to engage in self-deception leads to a loss in (ex-ante) expected utility:*

Corollary 1 *When information has purely hedonic value, the DM is strictly better off by not observing the signal: $E[u] > U(\Sigma)$. Furthermore, in order to observe the signal, the individual requires a "participation premium" of $q\psi_H(m_H^*) + (1 - q)\psi_L(m_L^*)$.*

Proposition 6 and Corollary 1 show that memory manipulation leads to the avoidance of information when individuals have preferences over their own attributes (i.e., they have "ego utility").

The most standard model of ego utility one could formulate consists of a basic application of expected utility theory. Let the space of possible attributes Θ be a non-empty subset of $\mathbb{R}$ and let $F(\cdot)$ denote the agent's prior distribution of θ . The DM has preferences that are represented by a strictly increasing von Neumann-Morgenstern utility function $u : \Theta \rightarrow \mathbb{R}$.

In this basic model, if the individual does not observe a signal that is informative about θ , her utility is $\int u(\theta) dF(\theta)$. If she observes a signal σ , the utility conditional on σ is $\int u(\theta) dF(\theta|\sigma)$. Hence, the expected utility of observing the signal is $\int_{\sigma} \int_{\theta} u(\theta) dF(\theta|\sigma) dG(\sigma)$, where G is the distribution of signals σ . By the law of iterated expectations, we have

$$\int u(\theta) dF(\theta) = \int_{\sigma} \int_{\theta} u(\theta) dF(\theta|\sigma) dG(\sigma),$$

so that *an individual with perfect memory and who behaves as an expected utility maximizer is always indifferent between observing the signal or not* when signals do not affect actions. In other words, in this standard model of ego utility, the fact that an individual has preferences over her expected attributes does not influence her decision of whether to acquire information. In particular, as in Blackwell's theorem, more information cannot hurt the individual.

Note that the result above holds regardless of the shape of the utility function u . In order to affect the decision of whether to acquire information, the utility function must be a non-linear function of *probabilities*. Several models of information acquisition have, thus, assumed that utility functions are non-linear in probabilities.²⁹ Our model also leads to a utility function that is non-linear in probabilities. However, the non-linearity arises endogenously through memory manipulation. Therefore, the present model can be seen as providing a cognitive foundation for a model of information acquisition.

Proposition 6 shows that the DM will prefer not to collect some information if its objective value V is lower than the expected costs from memory manipulation $q\psi_H(m_H^*(a^*)) + (1 - q)\psi_L(m_L^*(a^*))$. In particular, she will always prefer not to observe information that is informative about her attributes θ but does not affect her actions b . For example, people will prefer not to know the outcome of a medical exam if the value of information is not sufficiently high (e.g. if a detected disease is not treatable) and if the exam has a potentially large impact on the person's self-image. Dawson et al. (2006) present experimental evidence supporting this result.³⁰

An immediate consequence of avoiding information correlated with one's skills is the possible desirability of "self-handicapping" strategies such as under-preparing for an examination or getting too little sleep before a physical exercise (Berglas and Baumeister, 1993). Self-handicapping strategies reduce the informational content of the signal, and therefore, the model predicts that a person may engage in such strategies if the expected costs are not too high.

In several environments, competition allows for more precise information about one's abilities. Thus, individuals may display a "fear of competition" and prefer environments where

²⁹For example, Philipson and Posner (1995) and Caplin and Eliaz (2003) analyze the case of testing for sexually transmitted diseases, Köszegi (2003) considers a model of patient decision-making, Köszegi (2006) studies information acquisition and financial decisions, and Caplin and Leahy (2004) study strategic information transmission. With the exception of Philipson and Posner (1995), who do not provide a justification for the assumption of a utility function that is non-linear in probabilities, all these papers depart from the standard expected utility model by adopting the Psychological Expected Utility model.

³⁰Dunning (2005) obtained the same result in the domain of academic ability.

outcomes are not directly comparable to the outcomes from other people. More generally, the model predicts that in environments where information is correlated with one's attributes, individuals typically face a trade-off between the objective value of information and the costs of self-deception. Coarser information structures reduce the objective value of information but cause lower self-deception costs.

4.1 Regret Aversion

In this subsection, I study how the agent's utility from the lottery changes as a function of her prior distribution about her attributes. This allows us to show that the model developed in this paper provides a formalization for the (informal) theory of regret aversion based on self-evaluation proposed by Josephs et al. (1992).

The Theory of Regret Aversion based on Self-Perceptions The theory of choice based on regret aversion was simultaneously proposed by Bell (1982) and Loomes and Sugden (1982). According to this theory, agents base their decisions not only on expected payoffs but also on the payoffs that they would have obtained if they had made other decisions. Because agents anticipate feeling regret or delight over their choice, they take this into account when making a decision.

Josephs et al. (1992) argued that the feeling of regret arises from an individual's self-evaluation that follows an outcome.³¹ They suggested that people with worse self-perceptions are more severely harmed by negative outcomes than those with better self-perceptions. Therefore, *individuals with low self-image would be more likely to make choices that minimize the possibility of regret.*

According to this theory of regret aversion based on self-perception, the premium required to observe a signal σ_a that is informative about the DM's attributes should be decreasing in the favorableness of the agent's prior distribution (see Figure 8). Denote by $U(\Sigma_a)$ the expected utility of observing signal σ_a and, as in Proposition 6, let $E[u]$ denote the expected utility from not observing the signal. Then, the theory predicts that $E[u] - U(\Sigma_a)$ should be decreasing in the favorableness of the agent's prior distribution over her attributes.

The Model Since the theory presented by Josephs et al. (1992) considers only choices where no ex-post actions are taken, assume that B is a singleton. Moreover, since the ex-ante decision consists of selecting a gamble, we interpret ex-ante actions $a \in A$ as a choice between different possible lotteries and assume that these actions do not affect the DM's utility function. The only way in which ex-ante actions $a \in A$ affect the agent's utility is through the different distributions associated with each lottery. For simplicity, I consider either the forgetfulness model of Example 1 or the limited memory model of Example 2.

In order to determine how the agent's attitude toward information is affected by her prior, let κ be a parameter that indexes her prior distribution. A higher parameter κ leads to a more favorable prior in the sense of first-order stochastic dominance:

$$\kappa' > \kappa \implies F(\theta; \kappa') \leq F(\theta; \kappa), \quad (12)$$

for all $\theta \in \Theta$, with strict inequality for some θ .

³¹See also Larrick (1993) for a similar discussion.

Figure 8: Regret Aversion based on Self-Perceptions

Denote the gain from observing a high signal instead of a low signal by

$$\Delta u(\kappa, a) = \int u(\theta) dF(\theta|\sigma_a = H; \kappa) - \int u(\theta) dF(\theta|\sigma_a = L; \kappa).$$

The assumption that individuals with worse self-perceptions are more severely harmed by negative outcomes than those with better self-perceptions states can be stated as:

Assumption 3. $\Delta u(\kappa, a)$ is decreasing in κ for all $a \in A$.

Recall that $U(\Sigma_a)$ and $E[u]$ were defined as the expected utility of observing signal and the expected utility from not observing the signal, respectively. Then, the prediction of the theory of regret aversion based on self-perceptions can be stated as follows:

Conjecture 1 (Josephs et al., 1992) $E[u] - U(\Sigma_a)$ is positive and decreasing in κ , for all $a \in A$.

Next, I show that under Assumption 3, the model implies that Conjecture 1 is true. Since there are no ex-post actions in this model, the only benefit from memory manipulation is the change in the DM's self-perceptions Δu . Therefore, the amount of memory manipulation is increasing in the self-image gain from observing a high signal Δu . Because, under Assumption 3, $\Delta u(\kappa, a)$ is decreasing in κ , we obtain:

Proposition 7 (Regret Aversion) *Suppose Assumption 3 holds and consider either the forgetfulness model of Example 1 or the limited memory model of Example 2. For any $a \in A$, the premium required to observe the signal σ_a is decreasing (in the sense of strong set order) in the the decision-maker's prior over her attributes indexed by parameter κ .*

Therefore, the model provides a formalization of the theory of regret aversion based on self-perception proposed by Josephs et al. (1992).

4.2 Prior-Dependent Attitude Towards Information

Proposition 6 showed that the DM will seek information if its objective value is greater than the expected costs of self-deception. This result contrasts with Blackwell's theorem, which states that more information cannot be harmful. Alternatively, Caplin and Leahy (2001) have proposed the Psychological Expected Utility (PEU) model which generalizes the expected utility model to allow for different attitudes towards information.

Eliaz and Spiegel (2006) have criticized the PEU model by showing that it is inconsistent with certain situations where a DM's preference for information varies with her prior distribution. In one example, they describe a patient who prefers more accurate medical tests when she is relatively certain of being healthy, yet she avoids these tests when she is relatively certain of being ill. In another example, they describe a manager that asks for their employees' opinion only when he is sufficiently certain that the new information will not cause her to change her views much. They proved that such behaviors are inconsistent with the PEU model. As a result, Eliaz and Spiegel have suggested that one should drop the Bayesian updating assumption.

As the following example shows, the model presented in this paper is consistent with these two examples described by Eliaz and Spiegel. Therefore, unlike the PEU model, the self-deception model leads to prior-dependent attitudes toward information while retaining Bayesian updating.³²

Example 3 *An individual must choose whether or not to take some medical exam. Let $a = E$ denote the choice of taking the exam and $a = NE$ denote the choice of not taking it. The exam is informative about the individual's health θ and has outcome $\sigma = H$ if the individual is healthy and $\sigma = L$ if she is not. If the individual takes the exam, she can undertake medical treatment $B = \{T, NT\}$, where $b = T$ and $b = NT$ denote the cases where she does and does not undertake the treatment.*

The individual's payoff from being healthy is 25. If she takes the medical exam, the individual has a cost of 5. Thus, her payoff conditional on a high signal is 25 if $b = T$ and 20 if $b = NT$. The agent's expected payoff conditional on a low signal is $\gamma(q)$. Undertaking the treatment can reduce the effects from the disease, which increases her expected payoff to $\gamma(q) + 1$. In order to be consistent with Assumption 3, assume that $\gamma(q)$ is increasing so that Δu is decreasing in the DM's prior distribution over her skills (indexed by the probability of observing a high signal q). Let $\gamma(\frac{1}{2}) = 10$ and $\gamma(1) = 20$. If the DM does not take the exam, she obtains an expected payoff of $25q + \gamma(q)(1 - q)$.

For simplicity, let the memory system be given by the forgetfulness model of Example 1 and suppose that memory manipulation is binary: $m_L \in \{-\frac{1}{2}, 0\}$ with $\psi_L(-\frac{1}{2}) = 3$. The decision problem is depicted in Figure 9.

It is straightforward to show that the agent chooses $m_L = 0$ when q is close to 1. In this case, since the objective value of information is positive and the cost of self-deception is zero, the DM always chooses to take the exam. When $q = \frac{1}{2}$, however, the DM engages in self-deception ($m_L = -\frac{1}{2}$). It can be shown that the expected cost from memory manipulation is greater than the objective value of information so that the DM prefers not to take the exam. Thus, unlike Eliaz and Spiegel's result on the PEU model, the DM may prefer to take the

³²Epstein (2007) presents a model of anticipatory utility. In the special case of rank-dependent expected utility, they show that their model is also able to accommodate the behavior from Eliaz and Spiegel's examples.

Figure 9: Decision Problem in Example 3

exam when she is relatively certain of being healthy ($q \approx 1$) but prefer not to take the exam for intermediate values of q .

5 Lotteries Over Money

We propose that the consequences of each bet include, besides monetary payoffs, the credit or blame associated with the outcome. Psychic payoffs of satisfaction or embarrassment can result from self-evaluation or from an evaluation by others. (Heath and Tversky, 1991, pp. 7-8)

In Section 4, outcomes $\sigma \in \{H, L\}$ consisted of purely informative signals, which affected the DM's utility only through her beliefs about her own attributes θ . This section considers outcomes that affect the DM's utility not only by providing information about θ but also directly through monetary payments. I show that the model leads to a theory of ambiguity aversion based on self-deception. The DM may reject gambles with small but positive expected value. Moreover, the model is consistent with the competence hypothesis proposed by Heath and Tversky (1991).

In order to focus on the implications of the model for the DM's preferences over monetary lotteries, I take A and B to be singletons so that the agent does not take any actions. Therefore, as in the model of Subsection 3.4, information has purely hedonic value. However, in the case of monetary lotteries, outcomes also have a direct effect on the DM's payoff through monetary payments.

As described in Section 3, the outcome $\sigma \in \{L, H\}$ is interpreted as a monetary payment. For notational simplicity, I omit a and b from the DM's von Neumann-Morgenstern utility function. Therefore, in this section, the DM's utility function is denoted by $u(\theta, x)$, where $x \in \mathbb{R}$ denotes the amount of money that she has. If $H > L$, a high outcome not only provides favorable information about the agent's attributes θ but also leads to a higher monetary payment. This is the natural assumption since, in most cases, the outcome associated with higher monetary payments is also associated with better attributes. If $L > H$, a high outcome provides favorable information about θ but provides a lower payment. The results in this paper hold for any L and H .

For simplicity, I assume that the utility function is additively separable between characteristics and money:

$$u(\theta, x) = v(\theta) + \tau(x),$$

for a strictly increasing function $v : \Theta \rightarrow \mathbb{R}$ and a function $\tau : \mathbb{R} \rightarrow \mathbb{R}$. Appendix A analyzes the general case. Let v_s denote the expected payoff from attributes conditional on recollection $\hat{\sigma} \in \{H, L, \emptyset\}$.

Under additive separability, monetary payments can be factored out of self 1's memory manipulation choice. Given an outcome $\sigma = s \in \{H, L\}$, she maximizes:

$$(\eta_s + m_s) v_s + (1 - \eta_s - m_s) v_\emptyset + \tau(s) - \psi_s(m_s).$$

Therefore, self 1 chooses the same amount of memory manipulation as in the purely hedonic signals model analyzed in Subsection 3.4. Proposition 4 then implies that the DM will never choose to remember a low outcome or forget a high outcome:

Corollary 2 *Suppose that ψ_s is strictly convex, $s \in \{H, L\}$. Then, in any PBE, $m_H^* > 0 \geq m_L^*$. Furthermore,*

$$\begin{aligned} v_H - v_L &\geq \psi'_H(1 - \eta_H) \implies m_H^* = 1 - \eta_H, \quad m_L^* = 0, \quad \text{and} \\ v_H - v_L &< \psi'_H(1 - \eta_H) \implies 0 < m_H^* < 1 - \eta_H, \quad m_L^* < 0. \end{aligned}$$

From equation (6), the DM's ex-ante expected utility is:

$$U(\Sigma) = q[v_H + \tau(H) - \psi_H(m_H^*)] + (1 - q)[v_L + \tau(L) - \psi_L(m_L^*)]. \quad (13)$$

It consists of the sum of the expected payoff from attributes, the expected monetary payoffs, and the expected cost of memory manipulation. Denote by U^I the utility of a lottery with the same distribution over monetary outcomes as the one above but whose monetary outcomes are uninformative about θ . Then, the DM's ex-ante expected utility can be written as

$$U(\Sigma) = U^I - q\psi_H(m_H^*) - (1 - q)\psi_L(m_L^*). \quad (14)$$

Because the DM takes no actions after observing the outcome (i.e., B is a singleton), information has no objective value. Therefore, the model implies that the uninformative lottery is strictly preferred.

Remark 2 Consider the entrepreneurship model described in Subsection 3.1. The DM will choose to become an entrepreneur if the expected monetary payoffs are greater than the expected costs of self-deception:

$$q\tau(H) + (1-q)\tau(L) \geq q\psi_H(m_H^*) + (1-q)\psi_L(m_L^*).$$

Baron (1999) presents evidence that individuals who become entrepreneurs find it easier to admit past mistakes to themselves. In a static environment, our model may easily lead to this result. Suppose, for example, that individuals have heterogeneous concerns for self-image or that homogeneous individuals play different equilibria of the game. Then those with a lower concern for self image or those who play equilibria with lower amounts of self-deception are precisely the ones who benefit the most from becoming entrepreneurs. Alternatively, Section 6 will establish that the expected cost of self-deception converges to zero as experience grows. Therefore, it could be the case that entrepreneurs were not different from other individuals ex-ante, but, as they have gained experience, their cost of admitting past mistakes decreased.

Remark 3 Consider the used car model described in Subsection 3.1. The individual will purchase the car if the expected payoff gain from the purchase is greater than the expected costs of forgetting a bad outcome.

Remark 4 Under the additive separability assumption, it is immediate to extend Proposition 7 to the case of monetary lotteries. Let κ index the DM's prior distribution as defined in equation (12). As in Assumption 3, assume that $\Delta v(\kappa)$ is decreasing in κ and consider either the forgetfulness model of Example 1 or the limited memory model of Example 2. Then, the premium $U^I - U(\Sigma)$ is positive and decreasing (in the sense of strong set order) in κ .

5.1 Probability Weights

In this subsection, I will consider a non-expected utility representation, where the decision-maker's expected utility from observing the signal is expressed as a weighted average of the utility in each state of the world $\sigma \in \{L, H\}$. The representation consists of a weighting function $w : [0, 1] \rightarrow \mathbb{R}$ such that the utility from participating in the lottery is

$$U(\Sigma) = w(q) \times u_H + [1 - w(q)] \times u_L,$$

where $u_s \equiv \int u(\theta, s) dF(\theta|\sigma = s)$, $s \in \{H, L\}$. Clearly, the decision maker is an expected utility maximizer if $w(q) = q$. Although the model does not feature ambiguity in the sense of an imprecise distribution of probabilities, I will follow the literature on decision-making under ambiguity and say that an agent is *ambiguity averse* if $w(q) < q$.³³

Proposition 8 shows that the ex-ante preferences of the DM can be represented by a non-expected utility and that the DM always displays ambiguity aversion when the outcomes from the lottery are informative about her attributes:³⁴

³³If we identify "unambiguous" lotteries as those whose outcomes are uninformative about the DM's attributes and follow the approach in Epstein (1999), it follows that the DM is ambiguity averse if and only if $w(q) < q$.

³⁴Note that the model of monetary lotteries becomes a model of purely hedonic signals (Subsection 3.4) when $\tau(H) = \tau(L)$. Thus, when information has purely hedonic value, the DM's expected utility from observing the signal can be represented by

$$U(\Sigma) = w(q) u_H + [1 - w(q)] u_L,$$

where $w(q) = q - \frac{q\psi_H(m_H^*) + (1-q)\psi_L(m_L^*)}{u_H - u_L}$. Furthermore: $w(0) = 0$, $w(1) = 1$, and $w(q) < q$ for all $q \in (0, 1)$.

Proposition 8 (Representation) *The DM’s expected utility from the monetary lottery can be represented by*

$$U(\Sigma) = w(q) u_H + [1 - w(q)] u_L, \quad (15)$$

where

$$w(q) = q - \frac{q\psi_H(m_H^*) + (1 - q)\psi_L(m_L^*)}{u_H - u_L}. \quad (16)$$

Furthermore, $w(0) = 0$, $w(1) = 1$, and $w(q) < q$ for all $q \in (0, 1)$.

Remark 5 *Note that the representation from equation (15) is not separable between probabilities and the utility u_s . Since the departure from linear probability weights is caused by memory manipulation, individuals who engage in more memory manipulation have lower probability weights $w(q)$. Furthermore, because the amount of memory manipulation is increasing in the marginal utility from attributes, it follows that the deviation from linear weighting is itself a function of u_s .*

5.2 Discussion

The model presented here implies that ambiguity aversion is a consequence of the lottery outcomes being informative about the DM’s attributes. Several experimental papers have related ambiguity aversion with the lotteries’ being influenced by an individual’s skill or knowledge.³⁵ First, some experiments have contradicted the idea that ambiguity aversion is related to the imprecision of the probability distribution of the events as is usually argued. Budescu, Weinberg, and Wallsten (1988), for example, compared decisions based on numerically, graphically (the shaded area in a circle), and verbally expressed probabilities. Numerical descriptions of a probability are less vague than graphic descriptions which, in turn, are less vague than verbal descriptions. Thus, if agents had a preference for more precise distributions, they should rank events whose probabilities have a numerical description first, graphic descriptions second, and verbal descriptions last. However, unlike ambiguity aversion would predict, subjects were indifferent between these lotteries. Indeed, the authors could not reject that the agents behaved according to subjective expected utility theory and weighted events linearly.³⁶

Heath and Tversky argued that people’s preferences over ambiguous events arise from the anticipation of feeling knowledgeable or competent.³⁷ Their interpretation of the Ellsberg paradox is as follows:

People do not like to bet on the unknown box, we suggest, because there is information, namely the proportion of red and green balls in the box, that is knowable in principle but unknown to them. The presence of such data makes people feel less knowledgeable and less competent and reduces the attractiveness of the corresponding bet. (Heath and Tversky, 1991, pp. 8)

³⁵See Goodie and Young (2007) for a detailed discussion of this literature.

³⁶See also Budescu et al. (2002).

³⁷Subsection 5.4 defines Heath and Tversky’s “competence hypothesis” more precisely and also briefly reviews the empirical evidence related to it.

Fox and Tversky (1995, pp. 585) proposed that ambiguity is caused by comparative ignorance. They have argued that “ambiguity aversion is produced by a comparison with less ambiguous events or with more knowledgeable individuals.” As in Heath and Tversky’s (1991) competence hypothesis, this “comparative ignorance hypothesis” states that ambiguity aversion is driven by the feeling of incompetence. Similarly, Goodie (2003) proposed the *perceived control* hypothesis, according to which ambiguity aversion is generated by an agent’s belief that the distribution of outcomes is influenced by attributes such as knowledge or skill.³⁸

As will be shown in Section 6, it is straightforward to embed the model in a dynamic setting where the DM updates beliefs according to Bayes’ rule. Therefore, the model provides a tractable framework where individuals display ambiguity aversion and still follow Bayes’ rule. Under the self-perception reinterpretation of ambiguity aversion, the difficulties in characterizing an updating rule under ambiguity do not arise.³⁹

5.3 Small-Stakes Risk Aversion

This subsection considers lotteries with small monetary stakes. It is shown that memory manipulation leads the DM to exhibit “zeroth-order” risk aversion, which has important implications. Standard expected utility maximizers exhibit second-order risk aversion. An individual with second-order risk aversion always accepts small gambles with positive expected value. Then, if the agent has reasonable levels of risk aversion with respect to lotteries with small stakes, she must display unrealistically high levels of risk aversion with respect to lotteries with large stakes (Samuelson, 1963; Rabin, 2000).

Segal and Spivak (1990) show that an individual with first-order risk aversion rejects small gambles as long as the positive expected value is sufficiently small. Therefore, several nonexpected utility models that feature first-order risk aversion have been proposed. However, Safra and Segal (2008) show that the inability to simultaneously explain an agent’s risk aversion over lotteries with small stakes and lotteries with large stakes can be generalized to non-expected utility models.⁴⁰ In this subsection, I show that the model allows us to reconcile risk aversion with respect to small lotteries with sensible levels of risk aversion with respect to large lotteries.⁴¹

Consider the lottery described previously. The certainty equivalent of the lottery is defined by the monetary amount $CE \in \mathbb{R}$ that makes the agent indifferent between participating in the lottery or receiving CE for sure:

$$\int u(\theta, CE) dF(\theta) = qu_H(H) + (1 - q)u_L(L) - q\psi_H(m_H^*) - (1 - q)\psi_H(m_L^*). \quad (17)$$

The risk premium associated with a lottery is defined as the difference between the expected payment and the certainty equivalent: $\pi = qH + (1 - q)L - CE$.

³⁸There is a large experimental literature on the effect of perceived control on risk-taking (c.f., Chau and Phillips, 1995, or Horswill and McKenna, 1999).

³⁹See, for example, Hanany and Kilbanoff (2007).

⁴⁰The crucial assumption in Segal and Spivak (1990) is that decision-makers have a unique preference relation over final-wealth distributions. Hence, their framework does not include gain-losses models such as Prospect Theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992).

⁴¹Fudenberg and Levine (2007) present a dual-self model of dynamic consumption where the agent’s exercise of self-control may also lead to risk aversion over lotteries with small stakes.

Let $s \in \{H, L\}$ be a binary random variable such that $E[s] = qH + (1 - q)L = 0$. Consider a lottery that pays $x = \varepsilon s$, where $\varepsilon > 0$. A decision maker has risk preferences of second order if $\lim_{\varepsilon \rightarrow 0+} \pi(\varepsilon)/\varepsilon = 0$. She is first-order risk averse if $\lim_{\varepsilon \rightarrow 0+} \pi(\varepsilon)/\varepsilon > 0$ is finite. She is zeroth-order risk averse if $\lim_{\varepsilon \rightarrow 0+} \pi(\varepsilon)/\varepsilon = +\infty$.

Note that the monetary lottery converges to a model of purely hedonic signals as ε approaches zero. Then, as shown in Corollary 1, the DM demands a strictly positive participation premium in order to observe the signal. Hence, the certainty equivalent of the lottery converges to $CE(0) < 0$ and

$$\lim_{\varepsilon \rightarrow 0+} \frac{\pi(\varepsilon)}{\varepsilon} = - \lim_{\varepsilon \rightarrow 0+} \frac{CE(\varepsilon)}{\varepsilon} = +\infty.$$

Thus, the individual exhibits zeroth-order risk aversion. This result is established formally in the following proposition:

Proposition 9 (Zeroth-Order Risk Aversion) *In any PBE, the DM exhibits zeroth-order risk aversion.*

Since outcomes are informative about the DM's attributes, the DM engages in memory manipulation. Therefore, even when the monetary payoffs converge to zero, she still demands a strictly positive risk premium. Hence, the individual displays zeroth-order risk aversion and displays risk aversion for lotteries with small stakes. However, as shown in Corollary 1, when the expected monetary stakes are larger than the DM's participation premium, she will accept to participate in the lottery. Thus, as the following example shows, the DM may be risk averse over lotteries with small stakes without displaying an unreasonable degree of risk aversion over lotteries with large stakes:

Example 4 (Safra and Segal, 2008) *Suppose an agent rejects a lottery that pays either -100 or 105 with equal probability at all wealth levels below $300,000$. Safra and Segal show that all standard non-expected utility models imply that this agent cannot accept a lottery that pays $-5,000$ or $10,000,000$ with equal probability for some wealth level below $300,000$. The model in this paper, however, is consistent with this behavior. Indeed, I will show that the DM may even accept the second lottery for all wealth levels below $300,000$.*

Suppose both lotteries have the same informational content about the DM's attributes θ . For simplicity, take the forgetfulness model of Example 1 with binary manipulation efforts $m_L \in \{-\frac{1}{2}, 0\}$ and let $\tau(x) = x$ for all $x \in \mathbb{R}$. Suppose that $\frac{1}{3}(v_H - v_L) > \psi_L(-\frac{1}{2})$ so that self 1 engages in memory manipulation: $m_L^ = -m$. Then, the DM rejects the first lottery and accepts the second lottery for all wealth levels below $300,000$ if*

$$\begin{aligned} \frac{1}{2}(v_L + W - 100) + \frac{1}{2}(v_H + W + 105) - \frac{1}{2}\psi_L\left(-\frac{1}{2}\right) &< \frac{1}{2}v_L + \frac{1}{2}v_H + W, \text{ and} \\ \frac{1}{2}(v_L + W - 5000) + \frac{1}{2}(v_H + W + 10000000) - \frac{1}{2}\psi_L\left(-\frac{1}{2}\right) &> \frac{1}{2}v_L + \frac{1}{2}v_H + W \end{aligned}$$

for all $W \leq 300,000$. These conditions are satisfied if

$$5 < \psi_L\left(-\frac{1}{2}\right) < \min\left\{\frac{1}{3}(v_H - v_L); 995,000\right\}. \quad (18)$$

Therefore, when inequality (18) is satisfied, the DM accepts the first lottery and rejects the second lottery for all wealth levels below $300,000$.

Figure 10: Experimental Results on the Competence Hypothesis

5.4 The Competence Hypothesis

Consider two lotteries with the same distribution over monetary outcomes. In the first lottery, outcomes are informative about the decision-maker's skills or knowledge whereas in the second they are not. If the information about one's skills or knowledge is not useful (i.e., the objective value of information from the first lottery is zero) and the individual is an expected utility maximizer, she should be indifferent between these lotteries. Since one's attributes are ambiguous, an ambiguity averse individual should prefer the lottery whose outcomes are uninformative about her skills or knowledge.

Heath and Tversky (1991) have studied this choice in a series of experiments. They have shown that people prefer the skill- or knowledge-dependent lottery in contexts where they feel knowledgeable or competent but prefer the skill- or knowledge-independent lottery in ones where they consider themselves ignorant or uninformed. In one experiment, for example, subjects were asked to answer several questions. Subjects also revealed (in an incentive-compatible way) their expected probability of answering the questions correctly. Afterwards, they chose between betting on their answers or participating in a lottery with the same expected probability of winning. The proportion of people who chose to bet on their answers is presented in Figure 10.

If decision-makers are expected utility maximizers and the value of information is zero, they should be indifferent between these two lotteries. Since Prospect Theory does not distinguish between sources of uncertainty in the specification of probability weighting function, it also predicts that people should be indifferent between these two lotteries. Therefore, in both cases, the proportion of individuals who bet on the knowledge-based lottery should be roughly constant at 50% when the value of information is zero. If the value of information is positive, individuals who behave according to either Expected Utility Theory or Prospect theory should prefer the knowledge-based lottery. Hence, in this case, the proportion of

Figure 11: Simulated Results on the Competence Hypothesis

individuals who bet on the knowledge-based lottery should be constant at 100%.⁴²

Heath and Tversky found a remarkably different pattern. The proportion of people who preferred to bet on the knowledge-based lottery instead of a knowledge-independent lottery *with the same expected probability of winning* was increasing in the judged probability.⁴³ In situations where the expected probability of winning was small, people preferred to bet on the knowledge-independent lottery. On the other hand, when the expected probability of winning was large, individuals preferred to bet on the knowledge-based lottery. This result was labeled the Competence Hypothesis.

The following examples show that our model is consistent with the Competence Hypothesis:

Example 5 Consider the forgetfulness model of Example 1. For simplicity, let memory manipulation be a binary variable $m_L \in \{-\frac{2}{3}, 0\}$, with $\psi_L(-\frac{2}{3}) = \frac{1}{4}$ and $\psi_L(0) = 0$. Suppose that the DM does not face any ex-ante choice. However, in order to have a positive objective value of information, suppose that she chooses between a low b_L and a high b_H action ex-post.

Let $v_s(b)$ denote the expected payoff from attributes conditional on outcome $s \in \{H, L\}$ and take the following payoffs:

$$v_H(b_H) = 6, v_H(b_L) = 5, v_L(b_L) = 1, v_L(b_H) = 0, \tau(H) = 1, \tau(L) = 0.$$

In Appendix D, I show that the DM prefers the attribute-dependent lottery if $q > \frac{11}{23}$ and prefers the attribute-independent lottery if $q < \frac{11}{23}$.

⁴²Under the assumption that deviations from linear weighting are generated by the lottery being influenced by the DM's knowledge or skill (e.g., Heath and Tversky, 1991, and Fox and Tversky, 1995), this pattern is also inconsistent with standard probability weighting functions that overweight small probabilities and underweight large probabilities (Tversky and Kahneman, 1992, and Prelec, 1998).

⁴³A number of other experiments have confirmed the predictions of the competence hypothesis (c.f., Keppe and Weber, 1995; Taylor, 1995; Kilka and Weber, 2000; Chow and Sarin, 2001; Fox and Weber, 2002; Kuehberger and Perner, 2003; Di Mauro, 2008).

Example 6 Take the same parameters from the previous example but suppose that memory manipulation is a binary variable $m_L \in \{-\bar{m}, 0\}$, where the parameter $\bar{m}$ is distributed according to a c.d.f. Φ on $[\frac{1}{3}, 1]$. Suppose that $q \geq \frac{1}{4}$ so that the objective value of information is positive.⁴⁴ In Appendix D, I show that the proportion of people who prefer the attribute-dependent lottery is:

$$\begin{aligned} & 0 \text{ if } q \in \left[\frac{1}{4}, \frac{7}{19}\right], \\ & \Phi\left(\frac{3}{4} - \frac{1-2q}{1-q}\right) \text{ if } q \in \left(\frac{7}{19}, \frac{1}{2}\right], \\ & \Phi\left(\frac{3}{4}\right) \text{ if } q \in \left(\frac{1}{2}, 1\right]. \end{aligned}$$

Therefore, consistent with the Competence Hypothesis, this proportion is increasing in q . Figure 11 depicts the case where $\bar{m}$ is uniformly distributed.

6 Practice makes perfect: The Repeated Model

The previous sections considered a decision-maker who observes an outcome once and makes inferences about her attributes based on her recollection of this outcome. In several situations, however, individuals participate in this process repeatedly. A professional investor, for example, is constantly deciding which investment to undertake and receives feedback about the success or failure of these investments very frequently. It is often argued that the biases in decision-making that we observe in experimental settings would be severely attenuated as individuals gain experience. This section presents a repeated version of the general model described in Section 3 and shows that this is indeed the case in this model. More precisely, I show that the behavior of the DM converges to the one predicted by expected utility theory as the number of observed signals grows.

Consider a repeated version of the general model described in Section 3. For simplicity, I assume that A is a singleton so the DM only chooses actions after observing the signal.⁴⁵ In each period $n \in \{1, 2, 3, \dots, N\}$, an independent draw of the signal $\sigma_n \in \{H, L\}$ is made. Each signal σ_n is observed with probabilities $\Pr(\sigma = H|\theta)$ and $\Pr(\sigma = L|\theta)$, where θ is the agent's 'true' attributes. The parameter θ is not known. Instead, the DM has a prior $F(\theta)$ about its distribution. Hence, the prior over the distribution of a signal σ_n is

$$\Pr(\sigma_n = s) = \int \Pr(\sigma_n = s|\theta) dF(\theta) \quad s \in \{H, L\},$$

where the conditional probability $\Pr(\sigma = H|\theta)$ is strictly increasing in θ .

After observing $\sigma_n \in \{H, L\}$, the DM engages in memory manipulation m_L and m_H . She recollects a signal $\hat{\sigma}_n \in \{H, L, \emptyset\}$. A history at time n is a sequence of recollected signals and actions:

$$h^{n-1} = ((\hat{\sigma}_1, \dots, \hat{\sigma}_{n-1}), (b_1, \dots, b_{n-1})) \in H^{n-1},$$

⁴⁴If $q < \frac{1}{4}$, then $b(\hat{\sigma}) = b_L$ for all $\hat{\sigma}$. Therefore, since actions are not a function of recollections, the objective value of information is zero.

⁴⁵See Remark 6.

where $H^{n-1} \equiv \{\emptyset, L, H\}^{n-1} \times B^{n-1}$ is the set of possible histories. Note that, in this model, the DM can only manipulate the recollection of a signal in the time that the signal occurred. After the recollection has been registered into the agent's memory, she can no longer distort it.⁴⁶

As in the static game, the agent's choice is modeled through a different self acting each time information is forgotten. Thus, in each period, a stage-1 self chooses memory manipulations $(m_{H,n}, m_{L,n}) : H^{n-1} \rightarrow [-\eta_H, 1 - \eta_H] \times [-\eta_L, 1 - \eta_L]$ to maximize the discounted sum of payoffs from all future stage-games. The discount rate is $\delta \in [0, 1)$. Then, a stage-2 self applies Bayes' rule and chooses an action $b_n : \{\emptyset, L, H\} \times H^{n-1} \rightarrow B$. For notational clarity, I omit the arguments from the profiles of actions and manipulations.⁴⁷

Definition 2 A PBE of the game is a strategy profile (b^*, m_H^*, m_L^*) and posterior beliefs $\mu(\cdot|\cdot)$ such that:

1. $m_{s,n}^*$ maximizes

$$(\eta_s + m_s) \{E_\mu [u(\theta, b_n^*, s) | (s, b(s); h^{n-1})] + \delta V(s, b(s); h^{n-1})\} \\ + (1 - \eta_s - m_s) \{E_\mu [u(\theta, b_n^*, s) | (\emptyset, b(\emptyset); h^{n-1})] + \delta V(\emptyset, b(\emptyset); h^{n-1})\} - \psi_s(m_s)$$

with respect to m_s , $s \in \{H, L\}$.

2. $b_n^* \in \arg \max_{b \in B} \{E_\mu [u(\theta, b, s) | \hat{s}, h^{n-1}] + \delta V(\hat{s}, b; h^{n-1})\}$, for $s \in \{H, L\}$ and $\hat{s} \in \{H, L, \emptyset\}$.
3. $\mu(\cdot|h)$ is obtained by Bayes' rule if $\Pr(h | m_{L,n}^*, m_{H,n}^*) > 0$, for all $h \in H^n \cup \{\emptyset, L, H\} \times H^{n-1}$,
4. The continuation payoff V satisfies, for all $(\hat{s}, b; h^{n-1}) \in H^n$,

$$V(\hat{s}, b; h^{n-1}) = E_\mu \left\{ \sum_{z=n+1}^N \delta^{z-n} \left[u(\theta, b_z^*, \sigma_z) - \Pr(\sigma_z = H) \psi_H(m_{H,z}^*) - \Pr(\sigma_z = L) \psi_L(m_{L,z}^*) \right] \middle| (\hat{s}, b, h^{n-1}) \right\}.$$

I am interested in the PBE of the game when N is large for a fixed $\delta \in [0, 1)$. Let $\hat{\theta}_n(h^n)$ denote the Bayes estimator of θ given history h^n

$$\hat{\theta}_n(h^n) \equiv \int \theta dF(\theta | h^n).$$

Note that $F(\theta | h^n)$ is a function of m_H and m_L .

I assume that $\eta_H > 0$ and that there exists some $\bar{m} > -\eta_L$ with $\psi_L(\bar{m}) \geq \sup_\theta \{u(b, \theta, \sigma)\} - \inf_\theta \{u(b, \theta, \sigma)\}$ for all b, σ .⁴⁸ This assumption ensures that the DM never forgets a signal

⁴⁶This assumption captures the psychological finding that most information loss occurs soon after it is obtained. Nevertheless, it is clearly an extreme assumption. In general, forgetting rates seem to follow a power law (Anderson, 1995). Therefore, a large fraction of the information is lost right after learning, and over time, the rate of forgetting slows down.

⁴⁷Thus, we write b_n^* instead of $b_n^*(\hat{\sigma}_n, h^{n-1})$, and $m_{\sigma,n}^*$ instead of $m_{\sigma,n}^*(\sigma, h^{n-1})$.

⁴⁸This is satisfied, for example, if $\lim_{m_L \rightarrow -\eta_L} \psi(m_L) = +\infty$.

$\sigma_n \in \{H, L\}$ with probability 1.⁴⁹ The first issue is whether the Bayes estimator of θ is consistent. In other words, does the DM eventually learn her true attributes after observing a sufficiently large number of signals?

If memory manipulation were constant, the answer would be immediate because in this case, the recollections would be i.i.d., and hence Doob's Consistency theorem would imply that $\hat{\theta}_n(h^n)$ converges to θ . This is formally stated in the following lemma:

Lemma 1 *Suppose $m_{H,n}(h^{n-1}) = \tilde{m}_H$ and $m_{L,n}(h^{n-1}) = \tilde{m}_L$ for all h^{n-1}, n and let $N \rightarrow \infty$. Then $\hat{\theta}_n \rightarrow \theta$ for almost all histories.*

When memory manipulation is endogenous, however, it is not immediate that the DM eventually learns her true type. Although observed signals σ_n are i.i.d., memory manipulation leads to non-independent and non-identically distributed recollections $\hat{\sigma}_n$. However, because the agent knows the equilibrium strategies, she knows the probability of each signal conditional on the recollection. Therefore, intuitively, the agent correctly updates the recollections and eventually learns her true type regardless of how much manipulation effort she exerts.

The following result will be used in order to show that this intuition is correct:

Lemma 2 *For any fixed history h^n , $F(\theta|h^n; m_H, m_L)$ is increasing in m_H and decreasing in m_L .*

The lemma above implies that, conditional on reaching each history, the agent always prefers that she had forgotten high signals and remembered low signals. Because the agent is ultimately concerned about σ_n , $F(\theta|h^n; m_H, m_L)$ is not a function of m_H and m_L in all histories that do not contain any $\hat{\sigma}_n = \emptyset$. However, whenever the agent recollects $\hat{\sigma}_n = \emptyset$, she is always better off when she forgets high signals and remembers low signals (since it reduces the probability of arriving at $\hat{\sigma}_n = \emptyset$ after a low signal $\sigma_n = L$). Hence, $(\theta|h^n; -\eta_H, 1 - \eta_L)$ first-order stochastically dominates $(\theta|h^n; m_H, m_L)$ for all m_H, m_L .⁵⁰

A straightforward implication of Lemma 2 is that:

$$E[\theta|h^n; 1 - \eta_H, \bar{m}] \leq E[\theta|h^n; m_H, m_L] \leq E[\theta|h^n; -\eta_H, 1 - \eta_L], \quad (19)$$

for all m_H and $m_L \geq \bar{m}$ and all histories h^n . But because Lemma 1 implies that both extremes in the inequality (19) converge to θ , it thus follows that the term in the middle converges and has limit θ . This result is formally stated in the following proposition:⁵¹

⁴⁹Either one of these conditions are needed to ensure identification. If $\eta_H = 0$ and $m_L(h^n) = \eta_L$, then $m_H(h^n) = 0$ for all h^n would imply that $\hat{\sigma}_n = \emptyset$. In this case, the Bayesian posterior would be equal to the prior and, therefore, there is no hope for the Bayes estimator to be consistent. This assumption is not satisfied in the model of Example 3.4.1 ($\eta_H > 0$ is violated). However, it is straightforward to adjust the arguments from this section to establish the same results for that model.

⁵⁰The first-order dominance (FOSD) is for fixed h^n . Since the probability of each history is itself a function of m_L and m_H , it does not follow that there is unconditional FOSD.

⁵¹Note that the probability of occurrence of a history $\Pr(h^n)$ is a function of the sequence of memory manipulations m_H and m_L . Because $\eta_H > 0$ and $\psi_L(\bar{m}) \geq \sup_{\theta} \{u(b, \theta, \sigma)\} - \inf_{\theta} \{u(b, \theta, \sigma)\}$ for some $\bar{m} > -\eta_L$, the sets of histories with zero measure is the same for all relevant manipulation efforts: $m_H(h^n) \in [-\eta_H, 1 - \eta_H]$ and $m_L(h^n) \in [-\eta_L, \bar{m}]$. Therefore, we omit any explicit reference to m_L and m_H when considering almost sure convergence of $\hat{\theta}_n(h^n)$.

Proposition 10 (Consistency) *Let $N \rightarrow \infty$. Then, $\hat{\theta}_n \rightarrow \theta$ for almost all histories.*

Proposition 10 shows that, regardless of the memory manipulation employed by the DM, she eventually learns her true attributes θ . Thus, the benefit of memory manipulation converges to zero, and therefore, memory manipulation converges to zero as the number of observed signals increases:

Proposition 11 (No Manipulation in the Long Run) *Let $N \rightarrow \infty$. Then, $m_{H,n} \rightarrow 0$ and $m_{L,n} \rightarrow 0$ for almost all histories.*

Suppose signals are purely informative. As in Section 4, omit the signal σ from the agent's utility function. Define the optimal action $b^O(\theta) \in B$ as the one that maximizes the agent's utility when her attributes θ are known: $b^O(\theta) \in \arg \max_{b \in B} u(b, \theta)$. Proposition 11 implies that $b_n \rightarrow b^O$ for almost all histories. Therefore, in the limit, the DM chooses the same actions as an expected utility maximizer who knows θ .

Consider the case of monetary lotteries, and as in Section 5, omit the actions from the utility function. The DM's ex-ante utility from observing an additional signal converges to $qu(H, \theta) + (1 - q)u(L, \theta)$, which is the same utility of an expected utility maximizer when the attributes θ are known.

Therefore, when signals are observed frequently enough, agents will not engage in self-deception and their behavior will converge to the behavior of standard expected utility maximizers. This is consistent with the usual intuition that people do not exhibit ambiguity aversion over frequently observed events or that experts are subject to much less biases (e.g. List, 2003, List and Haigh, 2003).⁵²

Remark 6 *In the preceding analysis, we have assumed that A is a singleton so that the DM does not take actions that affect the distribution of the signals σ_a . This assumption simplifies the notation and the proofs. It is not important for our results as long as the first-order stochastic dominance assumption (equation 1) is retained. Thus, as long as all signals $a \in A$ are informative about θ , the DM eventually learns her true attributes and does not engage in memory manipulation.*

If the agent has the choice of not observing any signal (see, for example, Subsections 7.1 and 7.2), then she may choose never to obtain any information about θ . In that case, her expected attributes would not converge.

7 Applications

This section presents two applications of the model. The first application provides a self-deception model of the endowment effect. The second application provides a self-deception rationale for people taking sunk investments into consideration when making decisions.

⁵²List (2003) also showed that experienced traders of sports paraphernalia show smaller endowment effects for everyday goods used in lab studies than novice traders. This result is also consistent with the model above if the ability to trade sports paraphernalia is correlated with the ability to trade other goods.

7.1 The Endowment Effect

An individual that satisfies the axioms of expected utility theory does not display a difference between the maximum willingness to pay for a good and the minimum compensation demanded to sell the same good (willingness to accept) when income effects are small. However, several empirical works have documented a discrepancy between these values. An individual tends to value one good more when the good becomes part of that person's endowment. Thaler (1980) labeled this phenomenon an "endowment effect."

Kahneman, Knetsch, and Thaler (1990) argued that the endowment effect was caused by loss aversion.⁵³ This subsection proposes an alternative explanation for the endowment effect. The main idea is that, in most markets, trading requires certain skills or knowledge. At the very least, the parties must form an expectation of how much each good is worth. In more complex markets, they must also estimate the future prices of the goods (which determine the opportunity cost of trading). Therefore, as in the used car example (Subsection 3.1) the outcome from the trade reveals information about how skillful the person is.

As we have seen previously, an individual that cares about her self-image and is subject to imperfect memory will engage in an activity that reveals information about her skills only if the objective value of information is greater than the expected memory cost. Therefore, she may prefer not to trade if the price is only slightly above the expected value of the good.

The model is a special case of the general framework described in Section 3. Let $a \in \{T, NT\}$ denote the DM's choice of whether or not to trade an object. Let π denote the gain from trade (in monetary terms), which is unknown by the agent. Trading leads to an outcome $\sigma \in \{H, L\}$, which affects the DM's utility both directly through the gain from trade π_σ and indirectly because it is informative about the DM's skills θ . After observing the outcome σ , the DM engages in memory manipulation $m_L(T)$ and $m_H(T)$. As in Section 5, let the DM's preferences over skills θ and money x be represented by $v(\theta) + \tau(x)$ and, with no loss of generality, normalize the monetary payoff from not trading to zero ($\tau(0) = 0$).

Equation (13) implies that the DM will prefer to trade if the expected gain from trade is greater than the expected memory cost:

$$q\tau(\pi_H) + (1 - q)\tau(\pi_L) \geq q\psi_H(m_H^*(T)) + (1 - q)\psi_L(m_L^*(T)).$$

Therefore, we have the following result:

Proposition 12 (Endowment Effect) *There exist $\bar{\tau}_1 \geq \bar{\tau}_2 > 0$ such that:*

1. *the DM agrees to trade in any PBE if $E[\tau(\pi)] \geq \bar{\tau}_1$,*
2. *the DM refuses to trade in any PBE if $E[\tau(\pi)] \leq \bar{\tau}_2$,*
3. *there exist PBE where the DM agrees to trade and PBE where the DM refuses to trade if $\bar{\tau}_1 > E[\tau(\pi)] > \bar{\tau}_2$.*

In particular, a risk neutral individual will demand a strictly positive premium in order to trade. A risk averse individual will demand an even greater premium.

⁵³According to loss aversion, losses are weighed substantially more than gains. Then, the cost of losing a good is much higher than the benefit of winning a good.

A standard explanation based on ambiguity aversion would argue that the object initially owned by the DM has a less ambiguous distribution than the other object. Therefore, an ambiguity averse agent would not agree to trade if the expected gain from trade is not sufficiently high. Recall from Subsection 5.1 that the self-deception model relates the degree of ambiguity aversion with the DM's attributes. Thus, in the present model, the endowment effect is due to the self-evaluation that follows trade. Since the outcome of the trade is informative about the agent's skills or knowledge and therefore leads to costly self-deception, the DM may require some strictly positive premium in order to trade.

7.2 Sunk Cost Effects

The consequences of any single decision (...) can have implications about the utility of previous choices as well as determine future events or outcomes. This means that sunk costs may not be sunk psychologically but may enter into future decisions. (Staw, 1981, pp. 578)

Standard decision theory shows that only incremental costs and benefits should influence decisions. Historical costs, which have already been sunk, should be irrelevant. However, evidence suggests that people often take sunk costs into account when making decisions.⁵⁴ Genesove and Mayer (2001), for example, studied the Boston housing market. They have shown that when expected prices fall below a the original purchase price, sellers set an asking price that exceeds the asking price of other sellers by between 25 and 35 percent of the difference.

In a field experiment, Arkes and Blumer (1985) randomly selected sixty people to buy season tickets to the Ohio University Theater and divided them in three groups of twenty. Patrons in the first group paid the full price (\$15). Those in the second group received a \$2 discount, and people in the last group received a \$7 discount. Patrons in the first group attended significantly more than those in the discount groups.

This subsection shows that the self-deception model leads to sunk-cost effects. Psychologists have long argued that self-deception may be an important cause of why sunk costs affect choice. For example, Staw (1976) has shown that being personally responsible for an inefficient investment is an important factor in choosing to persist on it. Brockner et al. (1986) have documented that persisting on an inefficient allocation of resources is increased when subjects are told that outcomes reflected their “perceptual abilities and mathematical reasoning.”⁵⁵

Whether previous investments succeed or fail has important effects on the decision maker's self-views. Then, as the opening quote suggests, a past choice may be associated with not simply sunk *monetary* costs but also real *psychological* costs. Abandoning a project usually involves admitting that a wrong decision was made. Therefore, an individual revising her position in the project reveals information about her skills or knowledge. As shown in Section 4, the DM will prefer to avoid such information if the cost of making an uninformed decision is not high enough. But, in this case, some projects with negative expected value will not be terminated.

⁵⁴Sunk costs effects are also called “irrational escalation of commitment”, the “entrapment effect”, or “too much invested to quit”.

⁵⁵See Brockner (1992) for a review of the literature.

Figure 12: Timing of the model

The model is a special case of the general framework described in Section 3. As in Section 5, I assume the DM's utility function is additively separable over attributes θ and money x . For simplicity, I also assume that the DM is risk neutral so that $u(\theta, x) = v(\theta) + x$.

The timing of the model is presented in Figure 12. First, the DM chooses whether to invest in a project that costs $K > 0$ and gives a random monetary payoff of π . Let $a_0 \in \{I, NI\}$ denote the investment choice, where $a_0 = I$ if the DM undertakes the investment and $a_0 = NI$ if she does not. After the sunk investment was made, the DM can reevaluate the value of the project at zero cost. Let $a_1 = E$ denote the case where DM reevaluates the project and $a_1 = NE$ otherwise. Reevaluating the project leads to a (purely informative) signal $\sigma \in \{H, L\}$. A high signal is good news, both about the profitability of the project π and about the DM's skills θ .

After observing the signal σ , the DM may engage in memory manipulation m_L and m_H which leads to a recollection $\hat{\sigma} \in \{H, L, \emptyset\}$. Then, she chooses whether or not to abort the project. I write $b = A$ if the project is aborted and $b = C$ if it is continued. If the project is aborted, the DM obtains a monetary payoff of 0. If it is not aborted, the DM has an expected monetary payoff conditional on signal $s \in \{H, L\}$ of π_s .

I assume that the project is ex-ante efficient $E[\pi] > 0$.⁵⁶ As was shown in Proposition 6, the agent will prefer to observe the signal σ if the objective value of information, $V = -(1 - q)\pi_L > 0$, is greater than the expected manipulation costs, $q\psi_H(m_H^*) + (1 - q)\psi_L(m_L^*) > 0$. Hence, if the loss from not aborting after a low signal are “not too large,” the DM will prefer not to reevaluate the project:

Proposition 13 (Sunk Cost Effect) *There exist $\bar{\pi}_1 \leq \bar{\pi}_2 < 0$ such that:*

1. *the DM reevaluates the project in any PBE if $\pi_L \leq \bar{\pi}_1$,*
2. *the DM does not reevaluate the project in any PBE if $\pi_L \geq \bar{\pi}_2$, and*
3. *there exist PBE where the DM reevaluates the project and the DM doesn't reevaluate the project if $\bar{\pi}_1 < \pi_L < \bar{\pi}_2$.*

Since reevaluating one's previous decision is informative about the person's skills or knowledge, it leads to self-deception. Therefore, the DM will prefer not to reevaluate her initial choice if the monetary loss π_L from continuing an inefficient project is lower than the expected cost of memory manipulation. Note that the key feature of the model is *not* the psychological cost from failure itself. The individual will eventually find out whether the

⁵⁶If the project is ex-ante inefficient, $E[\pi] \leq 0$, the problem would be trivial since the DM would never invest.

project is successful or not. However, by not reevaluating a project, the individual avoids the psychological cost from *self-deception*.⁵⁷

8 Conclusion

This paper proposed a model of choice under risk based on imperfect memory and self-deception. The model provides a unified explanation for a number of biases in decision-making. It also leads to non-expected utility representation that is consistent with recent experimental evidence relating ambiguity aversion to an individual’s skills or knowledge.

The model can be enriched in several directions by incorporating strategic components. Principal-agent relationships seem like a natural application of the theory. Since the outcome of the relationship is typically informative about the agent’s skill or knowledge, principals may prefer to offer contracts that do not completely reveal the outcome to the agent. Therefore, firms may prefer not to condition wages on economy-wide shocks. Similarly, CEOs may be “rewarded for luck.”

Another interesting direction is in the field of incomplete contracts. Contracts may be incomplete due to the contracting parties’ preferences for avoiding information correlated with their skills or knowledge.⁵⁸ However, because parties understand the consequences of contract forms and post-contractual decisions, the allocation of rights may matter for the outcomes. Therefore, the general framework proposed here may provide a behavioral model for a theory of ownership based on incomplete contracts.

The model can also be embedded in a general equilibrium model. Since self-deception leads to endowment effects, the model may provide an explanation for the low volume of trades of uncertain assets occurring in equilibrium.⁵⁹

Finally, the model can lead to interesting predictions when θ is interpreted as a parameter of anticipatory utility. Because anticipatory utility typically leads to a first-order gain from memory manipulation but only second-order costs through suboptimal decision-making, individuals will forget negative news and remember positive news with probability above their natural rates. For example, a model of portfolio allocation where signals σ are informative about the profitability of a risky asset may provide an explanation for why most investors hold extremely underdiversified portfolios and overinvest in stocks issued by the their employing firm.

Appendix A Non-Separable Preferences

In Section 5, the DM’s preferences were assumed to be additively separable between attributes and money. In this section, I consider general utility functions. It turns out that a main feature in this general model is the degree of complementarity between attributes and money. As will be discussed later, since a DM is not as affected by monetary outcomes when she is uninformed about her attributes when attributes and money are complementary, complementarity can be interpreted

⁵⁷The argument above is related to agency explanations. For example, as argued by Li (2007), in environments with adverse selection, agents may prefer not to change their opinions if this publicly conceals bad news about their abilities. It is unclear, however, whether agency concerns would play an important role in contexts where the decisions are not publicly observed.

⁵⁸Mukerji (1998) showed that ambiguity aversion may lead to incomplete contracts. Tirole (2008) considered a model where thinking about contingencies is costly. In his model, contracts may be “too complete.”

⁵⁹See Billot et al. (2000) for a model based on ambiguity aversion.

as providing “psychological insurance.” Therefore, the DM may prefer a lottery whose outcomes are informative about her attributes if the complementarity effect is greater than the costs of self-deception. Moreover, the resulting probability weighting function may have an “inverted S-shape” as in Tversky and Kahneman (1992) and Prelec (1998).

Let $u_\sigma(s) \equiv \int u(\theta, s) dF(\theta|\sigma)$ denote the expected utility from a monetary amount equal to s conditional on outcome σ . As in Corollary 2, it can be shown that, in any PBE, $m_H^* > 0 \geq m_L^*$.⁶⁰

Define the degree of complementarity between θ and money by

$$\chi(H, L) \equiv u_H(H) + u_L(L) - u_L(H) - u_H(L). \quad (20)$$

Note that $\chi(H, L) \geq 0$ if u has increasing differences and $\chi(H, L) \leq 0$ if u has decreasing differences. The additively separable case presented in the text corresponds to the case where $\chi(H, L) = 0$. The ex-ante expected utility from the lottery is

$$\begin{aligned} U(\Sigma) &= q(\eta_H + m_H^*)u_H(H) + (1-q)(\eta_L + m_L^*)u_L(L) \\ &\quad + q(1-\eta_H - m_H^*)u_\varnothing(H) + (1-q)(1-\eta_L - m_L^*)u_\varnothing(L) - MC, \end{aligned}$$

where $MC = q\psi_H(m_H^*) + (1-q)\psi_H(m_L^*)$ is the expected memory cost. Then, long but tedious algebraic manipulations yield

$$U(\Sigma) = qu_H(H) + (1-q)u_L(L) + z\chi(H, L) - MC. \quad (21)$$

where $z = \frac{q(1-q)(1-\eta_L - m_L^*)(1-\eta_H - m_H^*)}{q(1-\eta_H - m_H^*) + (1-q)(1-\eta_L - m_L^*)} > 0$ and $MC = q\psi_H(m_H^*) + (1-q)\psi_H(m_L^*)$.

The utility of a monetary lottery can be decomposed in three terms: First, the expected utility $qu_H(H) + (1-q)u_L(L)$ of the lottery when memory is perfect. Second, the expected manipulation costs MC . These two effects are precisely the same as in the additively separable case (see equation 13). The third effect, which is not present when the utility is additively separable, is the degree of complementarity between attributes and money. When signals are forgotten, there is probability $\alpha(m_H^*, m_L^*)$ that a high signal was observed and the complementary probability that a low signal was observed. Thus, forgetting a signal can be seen as providing “psychological insurance” to the agent. This raises her expected utility if θ and money are complementary ($\chi > 0$) and decreases her expected utility if they are substitutes ($\chi < 0$).

Proceeding as in Subsection 5.1, it follows that the DM’s expected utility can be represented by

$$U(\Sigma) = w(q) \times u_H(H) + [1 - w(q)] \times u_L(L),$$

where $w(q) = q + \frac{z\chi(H, L) - MC}{u_H(H) - u_L(L)}$. Moreover, it is straightforward to show that $w(0) = 0$, and $w(1) = 1$. Therefore, when attributes and money are complementary, the DM may exhibit ambiguity loving behavior. In particular, the following example shows that the model may lead to an inverted S-shaped probability weighting function:

Example 7 (Inverted S-shaped Probability Weighting Function) *Consider the limited memory model of Example 2 and suppose that the manipulation effort is a binary variable: $m_H \in \{0, \frac{3}{4}\}$, where $\psi_H(\frac{3}{4}) = \frac{1}{5}$. Let $\chi(H, L) = u_H(H) - u_H(L) = 1$. Then, self 1 chooses to engage in memory manipulation if $q \in (0, \frac{11}{12})$. It is straightforward to show that, for values of q such that the DM engages in memory manipulation, the probability weighting function has an inverted S-shape:*

$$w(q) \begin{cases} > q & \text{if } q \in (0, \frac{1}{2}) \\ < q & \text{if } q \in (\frac{1}{2}, \frac{11}{12}) \end{cases}.$$

⁶⁰ As in Corollary 2, it can also be shown that

$$\begin{aligned} u_H(H) - u_L(L) &\geq \psi'_H(1 - \eta_H) \implies m_H^* = 1 - \eta_H, \quad m_L^* = 0, \text{ and} \\ u_H(H) - u_L(L) &< \psi'_H(1 - \eta_H) \implies 0 < m_H^* < 1 - \eta_H, \quad m_L^* < 0. \end{aligned}$$

As in Section 5, denote by U^I the utility of a lottery with the same distribution over monetary outcomes as the one above but whose monetary outcomes are uninformative about θ . Rearranging equation (21), we obtain

$$U(\Sigma) = U^I + y\chi(H, L) - MC, \quad (22)$$

where $y = q(1 - q) \left[1 + \frac{(1 - \eta_L - m_L^*)(1 - \eta_H - m_H^*)}{q(1 - \eta_H - m_H^*) + (1 - q)(1 - \eta_L - m_L^*)} \right] > 0$. Consider the choice between the lottery Σ and another lottery with the same distribution over monetary outcomes but whose monetary outcomes are uninformative about θ . Equation (22) implies that the DM will prefer lottery Σ if the degree of complementarity is high enough or if the expected memory cost is low enough: $y\chi(H, L) \geq MC$. Therefore, when attributes and money are complementary, the DM may prefer the attribute-dependent lottery.

However, when the monetary lottery is “small” (i.e., when the lottery pays $x = \varepsilon s$ for ε low), the complementarity effect vanishes. Since the memory cost converges to a strictly positive number as ε converges to zero, it follows that the certainty equivalent of the lottery converges to $CE(0) < 0$. Therefore,

$$\lim_{\varepsilon \rightarrow 0+} \frac{\pi(\varepsilon)}{\varepsilon} = - \lim_{\varepsilon \rightarrow 0+} \frac{CE(\varepsilon)}{\varepsilon} = +\infty$$

and, for any degree of complementarity between attributes and money, the DM always exhibits zeroth-order risk aversion. This is formally stated in the following proposition:

Proposition 14 *In any PBE, the DM exhibits zeroth-order risk aversion.*

It is interesting to contrast the general model with the a model from the following example where the DM does not face memory costs:

Example 8 (Exogenous Memory Model) *Take $\psi_s(m_s) = +\infty$ for all $m_s \neq 0$. Thus, the agent cannot engage in endogenous memory manipulation. Let $\eta_s < 1$ so that the agent forgets outcome with (exogenous) probabilities $1 - \eta_s > 0$. If $\eta_H > \eta_L$, memory is selective in the sense that good news is more likely to be remembered than bad news.*

When memory manipulation is endogenous (and differentiable at $m_s = 0$), the effect from memory manipulation always dominates the complementarity effects and the DM displays zeroth-order risk aversion. When memory manipulation is exogenous, the order of risk aversion is determined by the degree of complementarity between attributes and money. Note that for small ε , attributes and money are complementary if $u'_H(0) < u'_L(0)$ and substitutes if $u'_H(0) > u'_L(0)$.

Proposition 15 *In the exogenous memory model: (i) the DM is first-order risk averse if $u'_L(0) > u'_H(0)$; (ii) the DM is first-order risk seeking if $u'_L(0) < u'_H(0)$; and (iii) the DM has second-order risk preferences if $u'_L(0) = u'_H(0)$.*

Therefore, the DM may display risk preferences of first order when there are no manipulation costs. Unlike when memory manipulation is endogenous, the DM may be first-order risk seeking or have risk preferences of second order.

Appendix B Non-Bayesian Framework

Throughout the paper, I have maintained the assumption that the DM understands that she engages in memory manipulation and, thus, interprets her recollections according to Bayes' rule. Therefore, in the model presented in the text, individuals are sophisticated. In this section, I consider the case of naive individuals. As in Mullainathan (2002), naive individuals are unaware of their imperfect memory and interpret recollections as if they were the true outcomes. Two interesting

features arise under naiveté. First, unlike the model of sophisticated individuals, the equilibrium is unique. Second, decision makers may prefer to observe a signal even if it has no objective value. As a consequence, they may display ambiguity seeking behavior even under additive separability between attributes and money. Moreover, the individual may exhibit zeroth-order risk seeking behavior.

Consider a naive decision maker (NDM), who is unaware of her memory manipulation efforts. Therefore, she applies Bayes' rule as if her recollections were generated by the memory system when she does not engage in memory manipulation, i.e. $m_L = m_H = 0$. When $\hat{\sigma} \in \{H, L\}$, she correctly infers that outcome $\sigma = \hat{\sigma}$ has been observed in period 1. However, when an outcome is forgotten, she attributes weight

$$\rho \equiv \frac{q(1 - \eta_H)}{q(1 - \eta_H) + (1 - q)(1 - \eta_L)} \quad (23)$$

to a high outcome and $(1 - \rho)$ to a low outcome. I refer to such updating rule as *naive Bayes' rule*. The following definition proposes an adaptation of the PBE concept to naive decision makers:

Definition 3 *A Perfect Naively Bayesian Equilibrium (PNBE) of the game is a strategy profile $(a^*, b^*, m_H^*(a), m_L^*(a))$ and posterior beliefs $\mu(\cdot|\hat{\sigma}_a)$ such that:*

1. $a^* \in \arg \max_{a \in A} \left\{ \begin{array}{l} E_{\hat{\sigma}_a} [E_\mu [u(a, b_a^*(\hat{\sigma}_a), \theta, \sigma_a) | \hat{\sigma}_a] | m_L^*(a), m_H^*(a)] \\ - q\psi_H(m_H^*(a)) - (1 - q)\psi_L(m_L^*(a)) \end{array} \right\};$
2. $m_s^*(a) \in \arg \max_{m_s} \left\{ \begin{array}{l} (\eta_s + m_s) E_\mu [u(a, b_a^*(\hat{\sigma}_a), \theta, s) | \hat{\sigma}_a = s] \\ + (1 - \eta_s - m_s) E_\mu [u(a, b_a^*(\hat{\sigma}_a), \theta, s) | \hat{\sigma}_a = \emptyset] - \psi_s(m_s) \end{array} \right\},$
 $s \in \{H, L\};$
3. $b_a^*(\hat{\sigma}) \in \arg \max_{b \in B} \{E_\mu [u(a, b, \theta, \sigma_a) | \hat{\sigma}_a = \hat{\sigma}]\};$
4. $\mu(\theta | \hat{\sigma}_a = \hat{\sigma})$ is obtained by naive Bayes' rule if $\Pr(\hat{\sigma}_a = \hat{\sigma} | m_L = m_H = 0) > 0, \forall \hat{\sigma} \in \{L, H, \emptyset\}$.

Conditions 1 – 3 are the same as in the PBE concept. Condition 4 modifies the standard Bayesian condition by requiring agents to follow the naive Bayes rule instead.

An interesting special case of this naive framework is obtained when we take the forgetfulness memory system of Example 1. Recall that if the state $\emptyset$ is interpreted as a recollection of a high outcome, then the model from Example 1 becomes one where the agent is able to convince herself that a low outcome was a high outcome by engaging in memory manipulation. Suppose an individual recollects a high outcome (i.e., $\hat{\sigma} = \emptyset$). If this individual is sophisticated, she then corrects for her memory imperfection and attributes some (Bayesian) weight to the possibility that she has observed a low outcome but managed to convince herself that the outcome was high instead. On the other hand, a naive individual believes her recollection is correct and attributes full weight to a high outcome ($\rho = 1$).

B1 Equilibrium Uniqueness

This subsection establishes that a PNBE exists and, under mild conditions, is unique. The naive updating rule implies that the NDM's expected utility given $\hat{\sigma} = \emptyset$ is

$$u_\emptyset(a, b, \sigma) = \rho u_H(a, b, \sigma) + (1 - \rho) u_L(a, b, \sigma). \quad (24)$$

Upon observing an outcome $s \in \{H, L\}$, self 1 maximizes:

$$(\eta_s + m_s) u_s(a, b_a(s), s) + (1 - \eta_s - m_s) u_\emptyset(a, b_a(\emptyset), s) - \psi_s(m_s). \quad (25)$$

The key feature of the naive updating rule is that it is not a function of the amount of memory manipulation employed by self 1. This greatly simplifies the computation of the PNBE of the model since, unlike in the sophisticated case, there is no feedback between self 2's expectation of the manipulation exerted by self 1 and self 1's manipulation choice. Then, the equilibrium amount of manipulation is determined by the maximum of expression (25).

Proposition 16 *There exists a PNBE. Furthermore, if ψ_s is strictly convex and $b_a(\hat{s})$ is a (single-valued) function where $s \in \{H, L\}$ and $\hat{s} \in \{H, L, \emptyset\}$, the PNBE is essentially unique.⁶¹*

Proof. Existence follows the same argument as Proposition 1. Note that Condition 3 from Definition 3 implies that $b_a(\hat{s})$ is not a function of self 1's memory manipulation. Strict convexity of ψ_s implies that expression (25) is strictly concave in m_s . Then, the equilibrium amounts of memory manipulation m_L^* and m_H^* are unique. Condition 4 implies that beliefs must also coincide in all recollections such that $\Pr(\hat{\sigma}_a = \hat{\sigma} | m_L^* = m_H^* = 0) > 0$. ■

Corollary 3 *The PNBE is essentially unique when either: (i) $u(a, \cdot, \theta, \sigma_a) : B \rightarrow \mathbb{R}$ is a strictly concave function, or (ii) B is a singleton (i.e., the individual does not take ex-post actions).*

Remark 7 *Suppose that B is finite and fix the natural rates of remembering an outcome η_L and η_H . Since the set of utility functions $u : \Theta \times A \times B \times \mathbb{R} \rightarrow \mathbb{R}$ such that $\arg \max_{b \in B} \{E_\mu[u(a, b, \theta, \sigma_a) | \hat{\sigma}_a = \hat{\sigma}]\}$ contains more than one element is nowhere dense, it follows that the PNBE is essentially unique for generic utility functions when the set of ex-post actions B is finite.*

The generic uniqueness of the PNBE contrasts with the multiplicity of the PBE discussed in Subsubsection 3.4.2. Multiplicity arises from the fact that self 1 affects self 2's equilibrium inference when the individual is sophisticated. In the naive case, because there is not effect from memory manipulation on self 2's inference, uniqueness is obtained.

B2 Ambiguity-Seeking Behavior

For simplicity, consider the forgetfulness memory system of Example 1 and, as in the Section 5, assume that the utility is additively separable between attributes and money. Then, the equilibrium amount of memory manipulation is $m_L^* = \min \left\{ \psi_L'^{-1}(\Delta u); 1 \right\}$. The ex-ante expected utility of the NDM is

$$\begin{aligned} U(\Sigma) &= (1 - q)(1 + m_L^*)u_L + [q - (1 - q)m_L^*]u_\emptyset - (1 - q)\psi_L(m_L^*) \\ &= (1 - q)(1 + m_L^*)u_L + [q - (1 - q)m_L^*]u_H - (1 - q)\psi_L(m_L^*), \end{aligned}$$

where the second inequality uses the fact that $u_\emptyset = u_H$. The NDM prefers to observe the signal Σ if and only if the expected improvement in self-image $|m_L^*|\Delta u$ is greater than the cost of memory manipulation $\psi_L(m_L^*)$.⁶² Thus, naive individuals may prefer to observe signals even if the objective value of information (which in this case is zero) is lower than the expected costs of manipulation.

Remark 8 *Proceeding as in Proposition 8, it follows that the NDM's expected utility from the monetary lottery can be represented by*

$$U(\Sigma) = w(q)u_H + [1 - w(q)]u_L,$$

where $w(q) = q - (1 - q) \left[m_L^* + \frac{\psi_L(m_L^*)}{\Delta u} \right]$, $w(0) = 0$, and $w(1) = 1$. Thus, the NDM is ambiguity averse if $|m_L^*|\Delta u < \psi_L(m_L^*)$ and ambiguity seeking if the reverse inequality is satisfied. Hence, a naive individual may be ambiguity seeking even when the utility function is additively separable between attributes and money.

⁶¹The PNBE is essentially unique in the sense that, all PNBE feature the same choices of actions a and b , manipulation efforts m_L and m_H , and beliefs given recollections that are believed to be reached with positive probability (i.e., $(\hat{\sigma}_a = \hat{\sigma} | m_L = m_H = 0) > 0$). Equilibria may diverge only with respect to beliefs at recollections that are not believed to be reached with positive probability. Obviously, one can ensure uniqueness of beliefs in all recollections by assuming that the NDM believes that all recollections are reached with positive probability: $0 < \min\{\eta_H, \eta_L\} < 1$.

⁶²As in Subsection 4.1, the NDM's surplus from observing a signal is decreasing in the favorableness of her prior distribution over her attributes under Assumption 3. However, unlike Conjecture 1, this surplus may be positive when the individual is naive.

B3 Zeroth-Order Risk Seeking Behavior

This subsection shows that the NDM may be zeroth-order risk seeking. As in Subsection 5.3, consider a lottery that pays $x = \varepsilon s$, $s \in \{H, L\}$, where $qH + (1-q)L = 0$. Let $m_s^*(\varepsilon)$ denote the equilibrium amount of memory manipulation as a function of ε . The certainty equivalent of this lottery is:

$$\begin{aligned} \int u(\theta, CE(\varepsilon)) dF(\theta) &= (1-q)(1+m_L^*(\varepsilon))u_L(L) \\ &\quad + [q - (1-q)m_L^*(\varepsilon)]u_H(H) - (1-q)\psi_L(m_L^*(\varepsilon)) - q\psi_H(m_H^*(\varepsilon)), \end{aligned}$$

Recall that $u_\sigma(s) \equiv \int u(\theta, s) dF(\theta|\sigma) u_s(\sigma) = v_\sigma + \tau(s)$, where the last equality follows from additive separability. Then, taking the limit as $\varepsilon \rightarrow 0_+$, we obtain:

$$\tau(CE(0)) = -m_L^*(1-q)\Delta v - (1-q)\psi_L(m_L^*) - q\psi_H(m_H^*).$$

Hence, $CE(0) > 0$ if $|m_L^*(0)|(v_H - v_L) > \psi_L(m_L^*(0)) + \frac{q}{1-q}\psi_H(m_L^*(0))$ and the NDM is zeroth-order risk seeking. In the opposite case, the NDM is zeroth-order risk averse. Thus, we have established the following result:

Proposition 17 *The NDM is:*

- *zeroth-order risk averse if $|m_L^*(0)|(v_H - v_L) < \psi_L(m_L^*(0)) + \frac{q}{1-q}\psi_H(m_L^*(0))$, and*
- *zeroth-order risk seeking if $|m_L^*(0)|(v_H - v_L) > \psi_L(m_L^*(0)) + \frac{q}{1-q}\psi_H(m_L^*(0))$.*

Appendix C Finite Number of Realizations

In the main text, we assume that each outcome σ_a may be either high or low. It is straightforward to generalize this framework to allow for any finite number of possible outcomes. Suppose that, given action $a \in A$, an outcome $\sigma_a \in \{1, 2, \dots, S_a\}$ is realized, $S_a \geq 2$. Outcomes are ordered by first-order stochastic dominance:

$$F(\theta|\sigma_a = s) \leq F(\theta|\sigma_a = s+1)$$

for all $\theta \in \Theta$, $s \in \{1, 2, \dots, S_a\}$ and $a \in A$, with strict inequality for some value of θ .

An outcome $s \in \{1, 2, \dots, S_a\}$ is remembered with probability $\eta_{s,a} + m_s$, where $\eta_{s,a} \in [0, 1]$. Self 1 exerts memory manipulation $m_s \in [-\eta_{s,a}, 1 - \eta_{s,a}]$, which costs $\psi_s(m_s) \geq 0$. Then, self 2 observes a recollection of the outcome σ_a , which is denoted by $\tilde{\sigma}_a \in \{1, 2, \dots, S_a, \emptyset\}$ and takes an action $b \in B$.

Preferences are represented by a von Neumann-Morgenstern utility function $u : \Theta \times A \times B \times \mathbb{R} \rightarrow \mathbb{R}$ which is strictly increasing in θ . When $u(\theta, a, b, x) = u(\theta, a, b, y)$ for all $x, y \in \mathbb{R}$, the model has *purely informative signals*. If signals are purely informative and A and B are singletons, we say that they have a *purely hedonic value*. When $u(\theta, a, b, x) \neq u(\theta, a, b, y)$ for some $x, y \in \mathbb{R}$ we say that the model has *monetary signals*.

It is straightforward to generalize the results in the text to this framework. For the representation result of Proposition 8, however, one should note that probability weights are no longer unique when $S_a > 2$.

Entrepreneurship Example The performance $s \in \{S, F\}$ of an entrepreneur is affected by two independent variables: her attributes θ and the external conditions $r \in \{1, 2, 3, \dots, R\}$. Attributes and external conditions are substitutes for the entrepreneur's performance. Therefore, given her performance $s \in \{S, F\}$, more favorable external conditions r provide bad news about the agent's attributes (in the sense of first-order stochastic dominance). The individual always recollects her performance s , but may manipulate her memory in order to change the rate at which she remembers the external conditions r .

This situation is modeled as follows. Let $\sigma \in \{S, F\} \times \{1, \dots, R\}$ denote the outcome of the project. Outcomes are ordered by first-order stochastic dominance:

$$(S, 1) \succeq_{FOSD} (S, 2) \succeq_{FOSD} \dots \succeq_{FOSD} (S, R) \succeq_{FOSD} (F, 1) \succeq_{FOSD} \dots \succeq_{FOSD} (F, R),$$

where we write $x \succeq_{FOSD} y$ if x first-order stochastically dominates y . Given an outcome (s, r) , self 1 chooses the probability at which the external conditions r are forgotten by exerting manipulation effort $m_{s,r}$. Then, self 2 applies Bayes' rule to the recollections $(s, \hat{r})$, where $\hat{r} \in \{r, \emptyset\}$. The agent's payoff net of manipulation costs given a recollection $(s, \hat{r})$ is

$$E[v(\theta, s) | s, \hat{r}] + \tau(s),$$

where $s \in \{S, F\}$ and $\hat{r} \in \{1, 2, \dots, R, \emptyset\}$. Figure 13 presents the agent's decision tree.

Figure 13: Entrepreneur Example

It is straightforward to extend the results from the general framework to this environment. In particular, expected manipulation costs are always strictly positive. Therefore, the agent will require the expected monetary payoffs from starting a new company to be strictly higher than the payoff from the previous job in order to become an entrepreneur. Moreover, if all outcomes have the same natural rate of recollection (i.e., $\eta_{s,r} = \eta$ for all s, r), then $m_{s,1} \geq m_{s,2} \geq \dots \geq m_{s,R}$ with at least one strict inequality, $s \in \{S, F\}$. Hence, the agent will remember negative external conditions more frequently than positive ones.

Appendix D Proofs

D1 Proofs of Propositions and Lemmas

Proof of Proposition 1: Define $\hat{m}_L(m_L, m_H)$ and $\hat{m}_H(m_L, m_H)$ as the set of maxima of (4) and (5), respectively (these are the best-response correspondences of self 1). Since (4) and (5) are continuous and concave functions defined over a compact set, $\hat{m}_L(m_L, m_H)$ and $\hat{m}_H(m_L, m_H)$ are non-empty, convex, and compact sets. Define the transformation $T : [-\eta_L, 1 - \eta_L] \times [-\eta_H, 1 - \eta_H] \rightarrow [-\eta_L, 1 - \eta_L] \times [-\eta_H, 1 - \eta_H]$ by $T(m_L, m_H) = (\hat{m}_L(m_L, m_H), \hat{m}_H(m_L, m_H))$. Then, Kakutani's theorem establishes that there exists a fixed-point of T which is a PBE. ■

Proof of Proposition 2: The proof will use the following result:

Claim A1. $u_H(a, b_a(\emptyset), L) > u_L(a, b_a(L), L)$.

Proof of the claim. By revealed preference, $u_\emptyset(a, b_a(\emptyset), L) \geq u_\emptyset(a, b_a(L), L)$. From the definition of $u_\emptyset$,

$$\alpha u_H(a, b(\emptyset), L) + (1 - \alpha) u_L(a, b(\emptyset), L) \geq \alpha u_H(a, b(L), L) + (1 - \alpha) u_L(a, b(L), L),$$

where I omit the efforts m_H and m_L from $\alpha(m_H, m_L)$ for notational clarity. Rearranging, gives

$$\alpha [u_H(a, b_a(\emptyset), L) - u_H(a, b_a(L), L)] \geq (1 - \alpha) [u_L(a, b_a(L), L) - u_L(a, b_a(\emptyset), L)].$$

Then, revealed preference implies that

$$u_H(a, b_a(\emptyset), L) - u_H(a, b_a(L), L) \geq \frac{1 - \alpha}{\alpha} [u_L(a, b_a(L), L) - u_L(a, b_a(\emptyset), L)] \geq 0.$$

Thus, $u_H(a, b_a(\emptyset), L) \geq u_H(a, b_a(L), L)$. But, first-order stochastic dominance implies that $u_H(a, b_a(L), L) > u_L(a, b_a(L), L)$. Therefore, we have that $u_H(a, b_a(\emptyset), L) > u_L(a, b_a(L), L)$. ■

Proof of Proposition 2: Because $\eta_H < 1$, the set of strictly positive efforts given a high signal $(0, 1 - \eta_H]$ is nonempty. Since (5) is strictly concave, it suffices to show that its derivative evaluated at $m_H = 0$ is strictly positive:

$$u_H(a, b(H), H) - u_L(a, b(\emptyset), H) - \alpha(m_L^*(a), m_H^*(a)) [u_H(a, b(\emptyset), H) - u_L(a, b(\emptyset), H)] > 0, \quad (26)$$

for all $m_L^*(a), m_H^*(a)$, where I have used the fact that $\psi'_H(0) = 0$.

Note that, by revealed preference, $u_H(a, b(H), H) \geq u_H(a, b(\emptyset), H)$. Hence,

$$\frac{u_H(a, b(H)) - u_L(a, b(\emptyset))}{u_H(a, b(\emptyset)) - u_L(a, b(\emptyset))} \geq 1.$$

Rearranging, we obtain:

$$\begin{aligned} & u_H(a, b(H), H) - u_L(a, b(\emptyset), H) - \alpha(m_L^*(a), m_H^*(a)) [u_H(a, b(\emptyset), H) - u_L(a, b(\emptyset), H)] \\ & \geq u_H(a, b(H), H) - u_L(a, b(\emptyset), H) - [u_H(a, b(\emptyset), H) - u_L(a, b(\emptyset), H)] \geq 0. \end{aligned}$$

This shows that the expression on the left-hand side of (26) is non-negative. Suppose it is equal to zero. Then, by the previous inequality, it must be the case that $\alpha(m_L^*(a), m_H^*(a)) = 1$. But $\alpha(m_L^*(a), m_H^*(a)) = 1$ implies that $m_L^* = 1 - \eta_L$ which, from the Kuhn-Tucker condition of the maximization of (4), requires that

$$u_L(a, b_a(L), L) - u_H(a, b_a(\emptyset), L) \geq \psi'_L(1 - \eta_L) \geq 0.$$

But, from Claim A1, $u_L(a, b_a(L), L) - u_H(a, b_a(\emptyset), L) < 0$, which contradicts the inequality above. Hence, $m_H^*(a) > 0$ for all $a \in A$. ■

Proof of Proposition 3: Define the function W_s as the expected utility of self 1 conditional on $\sigma = s$:

$$W_s(m_L, m_H, a, \{b_a\}) = (1 - \eta_s - m_s) [\alpha(m_L, m_H) u_H(a, b_a(\emptyset), H) + [1 - \alpha(m_L, m_H)] u_L(a, b_a(\emptyset), L)] + (\eta_s + m_s) u_s(a, b_a(s), s) - \psi_s(m_s).$$

For notational clarity, I omit the term a from $m_s^*(a)$. In any PBE, m_s^* solves:

$$\max_{m_s} \left\{ (1 - \eta_s - m_s) [\alpha(m_L^*, m_H^*) u_H(a, b_a(\emptyset), H) + [1 - \alpha(m_L^*, m_H^*)] u_L(a, b_a(\emptyset), L)] + (\eta_s + m_s) u_s(a, b_a(s), s) - \psi_s(m_s) \right\}.$$

Therefore, the envelope theorem implies that

$$\left. \frac{\partial W_s}{\partial m_L} \right|_{\substack{m_L = m_L^* \\ m_H = m_H^*}} = (1 - \eta_s - m_s^*) [u_H(a, b_a(\emptyset), H) - u_L(a, b_a(\emptyset), L)] \frac{\partial \alpha(m_L^*, m_H^*)}{\partial m_s}. \quad (27)$$

The DM's ex-ante utility is equal to

$$\mathcal{U}(m_H, m_L, a, \{b_a\}) = q W_H(m_L, m_H, a, \{b_a\}) + (1 - q) W_L(m_L, m_H, a, \{b_a\})$$

Thus,

$$\left. \frac{\partial \mathcal{U}(m_H, m_L, a, \{b_a\})}{\partial m_s} \right|_{\substack{m_H = m_H^* \\ m_L = m_L^*}} = q \frac{\partial W_H(m_H^*(a), m_L^*(a), a)}{\partial m_s} + (1 - q) \frac{\partial W_L(m_H^*(a), m_L^*(a), a)}{\partial m_s}.$$

Since $\frac{\partial \alpha(m_H, m_L)}{\partial m_H} \leq 0 \leq \frac{\partial \alpha(m_H, m_L)}{\partial m_L}$ with at least one inequality being strict, equation (27) yields

$$\frac{\partial \mathcal{U}(m_H, m_L, a, \{b_a\})}{\partial m_H} < 0 < \frac{\partial \mathcal{U}(m_H, m_L, a, \{b_a\})}{\partial m_L}.$$

Then, the result follows from the concavity of $\mathcal{U}$. ■

Proof of Proposition 4: From Proposition 2, it follows that $m_H^*(a) > 0$. First, we establish that $m_L^*(a) \leq 0$. By the strict concavity of equation (4), it suffices to show that its derivative evaluated at $m_L = 0$ is weakly negative:

$$-\alpha(m_L^*(a), m_H^*(a)) (u_H - u_L) - \underbrace{\psi_L'(0)}_0 \leq 0,$$

which is true because $\alpha(m_L^*(a), m_H^*(a)) \geq 0$ and $u_H > u_L$.

From Kuhn-Tucker's conditions of the maximization of (4), there exists a PBE with $m_L^* = 0$ if and only if

$$\alpha(0, m_H^*(a)) (u_H - u_L) = 0,$$

for some $m_H^*(a)$ that maximizes (5) given $m_L^*(a) = 0$.

Because $u_H > u_L$, this is satisfied if and only if $\alpha(0, m_H^*(a)) = 0$. But $\alpha(0, m_H^*(a)) = 0$ implies that $m_H^* = 1 - \eta_H$. From Kuhn-Tucker's conditions of the maximization of (5), there exists a PBE with $m_H^* = 1 - \eta_H$ if and only if

$$(u_H - u_L) [1 - \alpha(m_L^*(a), m_H^*(a))] \geq \psi_H'(1 - \eta_H),$$

for some m_L^* that maximizes (4). Substituting $\alpha(m_L^*(a), m_H^*(a)) = 0$, it follows that $(u_H - u_L) \geq \psi'_H(1 - \eta_H)$. ■

Proof of Proposition 5: Existence follows from Proposition 2. For a fixed m_L^* , self 1 solves:

$$\max_{-1 \leq m_L \leq 0} u_L - m_L \alpha(m_L^*, 0) \Delta u - \psi_L(m_L).$$

The Kuhn-Tucker conditions are:

$$\begin{aligned} \alpha(m_L^*, 0) \Delta u &\geq -\psi'_L(-1) \implies m_L = -1, \\ \alpha(m_L^*, 0) \Delta u &\leq 0 \implies m_L = 0, \text{ and} \end{aligned}$$

$$0 < \alpha(m_L^*, 0) \Delta u < -\psi'_L(-1) \implies \alpha(m_L^*, 0) \Delta u = -\psi'_L(m_L).$$

In the PBE, $m_L = m_L^*$. Substituting $\alpha(m_L, 0) = \frac{q}{q - (1-q)m_L}$ and using the implicit function theorem, it follows that the unique PBE has manipulation efforts:

$$\begin{aligned} m_L^* &= -1 \text{ if } \Delta u \geq -\frac{\psi'_L(-1)}{q}, \text{ and} \\ \frac{q}{q - (1-q)m_L^*} \Delta u &= -\psi'_L(m_L^*) \text{ if } \Delta u < -\frac{\psi'_L(-1)}{q}. \end{aligned}$$

The first claim follows by inspection. Let the cost of manipulation be $\psi_L(m_L, \kappa)$, where κ parametrizes the marginal cost of memory manipulation: $\frac{\partial^2 \psi}{\partial m_L \partial \kappa} < 0$. Therefore, higher κ 's lead to a higher marginal cost of memory manipulation ($-m_L \geq 0$). Then, differentiation of the condition (8) and an inspection of the condition for boundary equilibria establishes the second and third claims. ■

Proof of Proposition 6: Follows from equations (10) and (11). ■

Proof of Proposition 7: First, consider the forgetfulness model. From Corollary 1,

$$E[u] - U(\Sigma_a) = (1 - q_a) \psi_L(m_L^*(a)) \geq 0.$$

Then, Proposition 5 implies that the amount of manipulation $|m_L^*|$ is increasing in Δu . Since, by Assumption 3, $\Delta u(\kappa, a)$ is decreasing in κ for any a , it follows that $E[u] - U(\Sigma_a)$ is decreasing in κ .

Consider the limited memory model. From Corollary 1,

$$E[u] - U(\Sigma_a) = q_a \psi_H(m_H^*(a)) \geq 0.$$

It can be shown that the set of equilibrium manipulations is increasing in the benefit of manipulation Δu (in the sense of strong set order). Then, Assumption 3 and the monotonicity of $\psi_H(m_H)$ in $m_H \geq 0$ imply that the set of equilibrium premia $\{E[u] - U(\Sigma_a)\}$ is decreasing in κ (in the sense of strong set order). ■

Proof of Proposition 8: The representation follows equations (13) and (15). Note that $q = 0$ implies that, in any PBE, $m_L^* = 0$. Thus, $w(0) = 0$. Similarly, in any PBE, $q = 1$ implies $m_H^* = 0$ and, therefore, $w(1) = 1$. ■

Proof of Proposition 9: This is a special case of Proposition 14. ■

Proof of Lemma 1: Note that in this case recollections are i.i.d. Then, in order to apply Doob's consistency theorem, we need to check that there exists a set $A \in \{\emptyset, L, H\}$ such that $\theta_1 \neq \theta_2 \implies \Pr_{\theta_1}(A) \neq \Pr_{\theta_2}(A)$. In each period, the probability of each recollection $\hat{\sigma}$ (which are i.i.d.) is

$$\begin{aligned}\Pr(\hat{\sigma} = H|\theta) &= \Pr(\sigma = H|\theta) \times \eta_H, \\ \Pr(\hat{\sigma} = L|\theta) &= [1 - \Pr(\sigma = H|\theta)] \eta_L, \\ \Pr(\hat{\sigma} = \emptyset|\theta) &= \Pr(\sigma = H|\theta)(1 - \eta_H) + [1 - \Pr(\sigma = H|\theta)](1 - \eta_L).\end{aligned}$$

Since $\Pr(\sigma = H|\theta)$ is strictly increasing in θ , it follows that $\theta_1 > \theta_2$ implies $\Pr_{\theta_1}(\hat{\sigma} = H) > \Pr_{\theta_2}(\hat{\sigma} = H)$ and $\Pr_{\theta_1}(\hat{\sigma} = L) < \Pr_{\theta_2}(\hat{\sigma} = L)$, which verifies the condition. ■

Proof of Lemma 2: To simplify the notation, consider the distribution of q instead of the distribution of θ . This is without loss of generality since $q = \Pr(\sigma = H|\theta)$ is strictly increasing in θ . With some abuse of notation, I will write $F(q|h^n)$ for the c.d.f. of q given history h^n .

Note that actions $b_n \in B$ are functions of the sequence of recollections $\{\hat{\sigma}_1, \dots, \hat{\sigma}_n\}$. Therefore, to simplify notation and with no loss of generality, I omit the actions $\{b_1, b_2, \dots, b_n\}$ from histories. Thus, with some abuse of notation, I will refer to a history as a sequence of recollections $h^n = \{\hat{\sigma}_1, \dots, \hat{\sigma}_n\}$ in all the proofs in the appendix.

Denote by $h^{n \setminus k}$ the history $\{\hat{\sigma}_1, \dots, \hat{\sigma}_{k-1}, \hat{\sigma}_{k+1}, \dots, \hat{\sigma}_n\}$. I will use the following result:

Claim A2. For any history h^n , we have:

$$F(q|h^{n \setminus k}, \hat{\sigma}_k = H) \leq F(q|h^{n \setminus k}, \hat{\sigma}_k = L).$$

This claim states that, for any history, a high signal is good news about q and a low signal is bad news about q in terms of first-order stochastic dominance.

Note that the p.d.f. conditional on h^n is

$$f(q|h^n) = \frac{\prod_{t:\sigma_t=\emptyset} [q(1 - \eta_H - m_{H,t}^*) + (1 - q)(1 - \eta_L - m_{L,t}^*)] \times \prod_{t:\sigma_t=H} q(\eta_H + m_{H,t}^*) \times \prod_{t:\sigma_t=L} (1 - q)(\eta_L + m_{L,t}^*) \times f(q)}{\int \left\{ \prod_{t:\sigma_t=\emptyset} [q(1 - \eta_H - m_{H,t}^*) + (1 - q)(1 - \eta_L - m_{L,t}^*)] \times \prod_{t:\sigma_t=H} q(\eta_H + m_{H,t}^*) \times \prod_{t:\sigma_t=L} (1 - q)(\eta_L + m_{L,t}^*) \times f(q) \right\} dq}.$$

Let $\#H$ denote the number of times that a signal $\hat{\sigma} = H$ was recollected: $\#\{t : \hat{\sigma}_t = H\}$. Similarly, define $\#L$ as $\#\{t : \hat{\sigma}_t = L\}$.⁶³ Then, after some algebraic manipulations, we can write:

$$f(q|h^n) = \frac{\prod_{t:\sigma_t=\emptyset} [q(1 - \eta_H - m_{H,t}^*) + (1 - q)(1 - \eta_L - m_{L,t}^*)] \times q^{\#H} \times (1 - q)^{\#L} \times f(q)}{\int \prod_{t:\sigma_t=\emptyset} [q(1 - \eta_H - m_{H,t}^*) + (1 - q)(1 - \eta_L - m_{L,t}^*)] \times q^{\#H} \times (1 - q)^{\#L} \times f(q) dq}.$$

Note that $f(q|h^n)$ is not a function of $m_{H,t}$ and $m_{L,t}$ for any history h^n such that $\hat{\sigma}_t \neq \emptyset$. This follows from the signals σ_t being i.i.d. and the fact that $\hat{\sigma}_t = \sigma_t$ when $\hat{\sigma}_t \neq \emptyset$. Integrating the equation above, we obtain

$$F(x|h^n) = \frac{\int_0^x \prod_{t:\sigma_t=\emptyset} [q(1 - \eta_H - m_{H,t}^*) + (1 - q)(1 - \eta_L - m_{L,t}^*)] \times q^{\#H} \times (1 - q)^{\#L} \times f(q) dq}{\int_0^1 \prod_{t:\sigma_t=\emptyset} [q(1 - \eta_H - m_{H,t}^*) + (1 - q)(1 - \eta_L - m_{L,t}^*)] \times q^{\#H} \times (1 - q)^{\#L} \times f(q) dq}.$$

⁶³ Obviously, $\#H$ and $\#L$ are functions of histories. We omit this dependence for notational clarity.

We are now ready to prove the Claim:

Proof of Claim A2. We have to show that

$$\frac{\int_0^x \prod_{t:\sigma_t=\emptyset} [q(1-\eta_H-m_{H,t}^*) + (1-q)(1-\eta_L-m_{L,t}^*)] \times q^{\#H} \times (1-q)^{\#L} \times f(q) dq}{\int_0^1 \prod_{t:\sigma_t=\emptyset} [q(1-\eta_H-m_{H,t}^*) + (1-q)(1-\eta_L-m_{L,t}^*)] \times q^{\#H} \times (1-q)^{\#L} \times f(q) dq} \leq \frac{\int_0^x \prod_{t:\sigma_t=\emptyset} [q(1-\eta_H-m_{H,t}^*) + (1-q)(1-\eta_L-m_{L,t}^*)] \times q^{\#H-1} \times (1-q)^{\#L+1} \times f(q) dq}{\int_0^1 \prod_{t:\sigma_t=\emptyset} [q(1-\eta_H-m_{H,t}^*) + (1-q)(1-\eta_L-m_{L,t}^*)] \times q^{\#H-1} \times (1-q)^{\#L+1} \times f(q) dq}.$$

When $x = 0$, both sides become 0 and, when $x = 1$, both sides are equal to 1.

The derivative of the LHS with respect to x is

$$\frac{\prod_{t:\sigma_t=\emptyset} [x(1-\eta_H-m_{H,t}^*) + (1-x)(1-\eta_L-m_{L,t}^*)] \times x^{\#H} \times (1-x)^{\#L} \times f(x)}{\int_0^1 \prod_{t:\sigma_t=\emptyset} [q(1-\eta_H-m_{H,t}^*) + (1-q)(1-\eta_L-m_{L,t}^*)] \times q^{\#H} \times (1-q)^{\#L} \times f(q) dq},$$

and the derivative of the RHS with respect to x is

$$\frac{\prod_{t:\sigma_t=\emptyset} [x(1-\eta_H-m_{H,t}^*) + (1-x)(1-\eta_L-m_{L,t}^*)] \times x^{\#H-1} \times (1-x)^{\#L+1} \times f(x)}{\int_0^1 \prod_{t:\sigma_t=\emptyset} [q(1-\eta_H-m_{H,t}^*) + (1-q)(1-\eta_L-m_{L,t}^*)] \times q^{\#H-1} \times (1-q)^{\#L+1} \times f(q) dq}.$$

Note that $\frac{dRHS}{dq} > \frac{dLHS}{dq}$ if and only if

$$\begin{aligned} & \frac{\prod_{t:\sigma_t=\emptyset} [x(1-\eta_H-m_{H,t}^*) + (1-x)(1-\eta_L-m_{L,t}^*)] \times x^{\#H} \times (1-x)^{\#L} \times f(x)}{\int_0^1 \prod_{t:\sigma_t=\emptyset} [q(1-\eta_H-m_{H,t}^*) + (1-q)(1-\eta_L-m_{L,t}^*)] \times q^{\#H} \times (1-q)^{\#L} \times f(q) dq} \\ & \leq \frac{\prod_{t:\sigma_t=\emptyset} [x(1-\eta_H-m_{H,t}^*) + (1-x)(1-\eta_L-m_{L,t}^*)] \times x^{\#H-1} \times (1-x)^{\#L+1} \times f(x)}{\int_0^1 \prod_{t:\sigma_t=\emptyset} [q(1-\eta_H-m_{H,t}^*) + (1-q)(1-\eta_L-m_{L,t}^*)] \times q^{\#H-1} \times (1-q)^{\#L+1} \times f(q) dq}. \end{aligned}$$

Rearranging, we obtain:

$$\begin{aligned} & \frac{\int_0^x \prod_{t:\sigma_t=\emptyset} [q(1-\eta_H-m_{H,t}^*) + (1-q)(1-\eta_L-m_{L,t}^*)] \times q^{\#H} \times (1-q)^{\#L} \times f(q) dq}{\int_0^1 \prod_{t:\sigma_t=\emptyset} [q(1-\eta_H-m_{H,t}^*) + (1-q)(1-\eta_L-m_{L,t}^*)] \times q^{\#H-1} \times (1-q)^{\#L+1} \times f(q) dq} \\ & \leq \frac{\int_0^x \prod_{t:\sigma_t=\emptyset} [q(1-\eta_H-m_{H,t}^*) + (1-q)(1-\eta_L-m_{L,t}^*)] \times q^{\#H} \times (1-q)^{\#L} \times f(q) dq}{\int_0^1 \prod_{t:\sigma_t=\emptyset} [q(1-\eta_H-m_{H,t}^*) + (1-q)(1-\eta_L-m_{L,t}^*)] \times q^{\#H-1} \times (1-q)^{\#L+1} \times f(q) dq}. \end{aligned}$$

Thus, $\frac{dRHS}{dq} > \frac{dLHS}{dq}$ if and only if

$$\rho(x) > \frac{\int_0^1 \prod_{t:\sigma_t=\emptyset} [q(1-\eta_H-m_{H,t}^*) + (1-q)(1-\eta_L-m_{L,t}^*)] \times q^{\#H} \times (1-q)^{\#L} \times f(q) dq}{\int_0^1 \prod_{t:\sigma_t=\emptyset} [q(1-\eta_H-m_{H,t}^*) + (1-q)(1-\eta_L-m_{L,t}^*)] \times q^{\#H-1} \times (1-q)^{\#L+1} \times f(q) dq},$$

where $\rho(x) = \frac{x}{1-x}$. Since $\rho(0) = 0$, $\rho(1) = +\infty$, $\rho(x)$ is strictly increasing in x , and the term on the right is a positive constant, there exists a unique $\bar{x}$ such that

$$\rho(x) > (<) \frac{\int_0^1 \prod_{t:\sigma_t=\emptyset} [q(1-\eta_H - m_{H,t}^*) + (1-q)(1-\eta_L - m_{L,t}^*)] \times q^{\#H} \times (1-q)^{\#L} \times f(q) dq}{\int_0^1 \prod_{t:\sigma_t=\emptyset} [q(1-\eta_H - m_{H,t}^*) + (1-q)(1-\eta_L - m_{L,t}^*)] \times q^{\#H-1} \times (1-q)^{\#L+1} \times f(q) dq},$$

if $x < (>) \bar{x}$.

Therefore, we have that $\frac{dRHS}{dq} > \frac{dLHS}{dq}$ if $x < \bar{x}$ and $\frac{dRHS}{dq} < \frac{dLHS}{dq}$ if $x > \bar{x}$. Thus, the inequality is satisfied for all q (it is satisfied with strict inequality whenever $q \in (0, 1)$ and with equality at $q \in \{0, 1\}$). ■

Now, we are ready to prove the lemma:

Proof of Lemma 2. As shown previously, $F(x|h^n)$ is not a function of $m_{L,k}^*$ and $m_{H,k}^*$ for k such that $\hat{\sigma}_k \neq \emptyset$. Therefore, we only need to establish the results for k such that $\hat{\sigma}_k = \emptyset$.

Consider an arbitrary k such that $\hat{\sigma}_k = \emptyset$. Then, $F(x|h^n)$ is equal to

$$\begin{aligned} & \left(1 - \eta_H - m_{H,k}^*\right) \int_0^x q^{\#H+1} \times (1-q)^{\#L} \prod_{t \neq k: \sigma_t = \emptyset} [q(1-\eta_H - m_{H,t}^*) + (1-q)(1-\eta_L - m_{L,t}^*)] f(q) dq \\ & + \left(1 - \eta_L - m_{L,k}^*\right) \int_0^x q^{\#H} \times (1-q)^{\#L+1} \prod_{t \neq k: \sigma_t = \emptyset} [q(1-\eta_H - m_{H,t}^*) + (1-q)(1-\eta_L - m_{L,t}^*)] f(q) dq \\ & \frac{\left(1 - \eta_H - m_{H,k}^*\right) \int_0^1 q^{\#H+1} (1-q)^{\#L} \prod_{t \neq k: \sigma_t = \emptyset} [q(1-\eta_H - m_{H,t}^*) + (1-q)(1-\eta_L - m_{L,t}^*)] f(q) dq}{\left(1 - \eta_H - m_{H,k}^*\right) \int_0^1 q^{\#H+1} (1-q)^{\#L} \prod_{t \neq k: \sigma_t = \emptyset} [q(1-\eta_H - m_{H,t}^*) + (1-q)(1-\eta_L - m_{L,t}^*)] f(q) dq} \\ & + \left(1 - \eta_L - m_{L,k}^*\right) \int_0^1 q^{\#H} \times (1-q)^{\#L+1} \prod_{t \neq k: \sigma_t = \emptyset} [q(1-\eta_H - m_{H,t}^*) + (1-q)(1-\eta_L - m_{L,t}^*)] f(q) dq \end{aligned}$$

With some algebraic manipulations, it follows that $\frac{dF}{dm_{H,k}}(x|h^n) > 0$ if and only if

$$\begin{aligned} & \int_0^x q^{\#H} \times (1-q)^{\#L+1} \prod_{t \neq k: \sigma_t = \emptyset} [q(1-\eta_H - m_{H,t}^*) + (1-q)(1-\eta_L - m_{L,t}^*)] f(q) dq \\ & \frac{\int_0^1 q^{\#H} \times (1-q)^{\#L+1} \prod_{t \neq k: \sigma_t = \emptyset} [q(1-\eta_H - m_{H,t}^*) + (1-q)(1-\eta_L - m_{L,t}^*)] f(q) dq}{\int_0^x q^{\#H+1} \times (1-q)^{\#L} \prod_{t \neq k: \sigma_t = \emptyset} [q(1-\eta_H - m_{H,t}^*) + (1-q)(1-\eta_L - m_{L,t}^*)] f(q) dq} > \\ & \frac{\int_0^1 q^{\#H+1} \times (1-q)^{\#L} \prod_{t \neq k: \sigma_t = \emptyset} [q(1-\eta_H - m_{H,t}^*) + (1-q)(1-\eta_L - m_{L,t}^*)] f(q) dq}{\int_0^1 q^{\#H+1} \times (1-q)^{\#L} \prod_{t \neq k: \sigma_t = \emptyset} [q(1-\eta_H - m_{H,t}^*) + (1-q)(1-\eta_L - m_{L,t}^*)] f(q) dq} \end{aligned}$$

Note that the left-hand side is equal to $F(x|h^n, \hat{\sigma}_{n+1} = L)$, whereas the right-hand side is equal to $F(x|h^n, \hat{\sigma}_{n+1} = H)$. From the previous claim, it follows that $F(x|h^n, \hat{\sigma}_{n+1} = L) \geq F(x|h^n, \hat{\sigma}_{n+1} = H)$, which proves that the condition above is satisfied. Therefore, we have shown that $\frac{dF}{dm_{H,k}^*}(x|h^n) > 0$.

The argument for $\frac{dF}{dm_{L,k}^*}(x|h^n) < 0$ is analogous. ■

Proof of Proposition 10: The result is immediate from inequality 19, Lemma 1, and the fact that the sets of histories with zero measure are the same for all relevant manipulation efforts. ■

Proof of Proposition 11: In period N , conditions 1 and 2 from the definition of a PBE state that

$$\begin{aligned} m_{L,N}(L, h^{N-1}) & \in \arg \max_{m_L} (\eta_L + m_L) \int u(\theta) dF(\theta|L, h^{N-1}) \\ & + (1 - \eta_L - m_L) \int u(\theta) dF(\theta|\emptyset, h^{N-1}) - \psi_L(m_L), \end{aligned}$$

and

$$m_{L,N}(H, h^{N-1}) \in \arg \max_{m_H} \left\{ \begin{aligned} & (\eta_H + m_H) [\int u(\theta) dF(\theta|H, h^{N-1})] \\ & + (1 - \eta_H - m_H) [\int u(\theta) dF(\theta|\emptyset, h^{N-1})] - \psi_H(m_H) \end{aligned} \right\}.$$

From Proposition 10, it follows that $\int u(\theta) dF(\theta|h^N)$ converges to $u(\theta)$ for almost all histories. But, when $\int u(\theta) dF(\theta|h^N) = u(\theta)$, it follows that $m_L(L, h^{N-1})$ maximizes

$$(\eta_L + m_L) u(\theta) + (1 - \eta_L - m_L) u(\theta) - \psi_L(m_L) = u(\theta) - \psi_L(m_L),$$

which has a global maximum at $m_L = 0$. Hence, by continuity, it follows that $m_L(L, h^{N-1}) \rightarrow 0$ (a.s.). Similarly, when $\int u(\theta) dF(\theta|h^N) = u(\theta)$, $m_H(H, h^{N-1})$ maximizes $u(\theta) - \psi_H(m_H)$ so that $m_H(H, h^{N-1}) \rightarrow 0$ (a.s.). ■

Proof of Proposition 12: Given $\sigma = s$, self 1 maximizes

$$(\eta_s + m_s)(\theta_s + \pi_s) + (1 - \eta_s - m_s)[\alpha^* \theta_H + (1 - \alpha^*) \theta_L + \pi_s] - \psi_s(m_s),$$

where $\alpha^* = \alpha(m_L^*, m_H^*)$. Simplifying, this expression becomes:

$$(\eta_s + m_s) \theta_s + (1 - \eta_s - m_s)[\alpha^* \theta_H + (1 - \alpha^*) \theta_L] + \pi_s - \psi_s(m_s).$$

Therefore, the solution of the maximization program of self 1 is independent of π_s . It thus follows the set of manipulation efforts $m_s^*(T)$ that are part of a PBE is the same for all π_L and π_H , $s \in \{H, L\}$.

The self 1 chooses $a = T$ if

$$q\pi_H + (1 - q)\pi_L \geq q\psi_H(m_H^*(T)) + (1 - q)\psi_L(m_L^*(T)).$$

The result then follows from the fact that the left-hand side is not a function of π_H and π_L . ■

Proof of Proposition 13: The expected utility of self 1 if she chooses (I, NE) is $q(\theta_H + \pi_H) + (1 - q)(\theta_L + \pi_L)$. Her expected utility if she chooses NI is $q\theta_H + (1 - q)\theta_L$. Because $q\pi_H + (1 - q)\pi_L > 0$, it follows that NI is never chosen.

If self 1 chooses (I, E) , she obtains:

$$q(\theta_H + \pi_H) + (1 - q)\theta_L - q\psi_H(m_H^*) - (1 - q)\psi_L(m_L^*).$$

Therefore, (I, E) is chosen if

$$q(\theta_H + \pi_H) + (1 - q)\theta_L - q\psi_H(m_H^*) - (1 - q)\psi_L(m_L^*) \geq q(\theta_H + \pi_H) + (1 - q)(\theta_L + \pi_L).$$

Rearranging, we obtain

$$-(1 - q)\pi_L \geq q\psi_H(m_H^*) + (1 - q)\psi_L(m_L^*). \quad (28)$$

Proceeding as in the proof of Proposition 12, it can be shown that the set of manipulation efforts $m_s^*(I, E)$ that are part of a PBE is the same for all π_L and π_H , $s \in \{H, L\}$. Then, the result follows immediately from equation (28). ■

Proof of Proposition 14: For any PBE, define the expected manipulation cost as $MC(\varepsilon) \equiv q\psi_H(m_H^*(\varepsilon)) + (1 - q)\psi_L(m_L^*(\varepsilon))$. Note that $\lim_{\varepsilon \rightarrow 0^+} \chi(\varepsilon H, \varepsilon L) = 0$. Therefore, for small ε , equation (17) becomes:

$$\int u(\theta, CE(\varepsilon)) dF(\theta) = qu_H(\varepsilon H) + (1 - q)u_L(\varepsilon L) - MC(\varepsilon). \quad (29)$$

Since $MC(0) > 0$ and, by the Theorem of the Maximum, $MC(\varepsilon)$ is continuous, it follows that $MC(\varepsilon) > 0$ for small ε . Hence, $\lim_{\varepsilon \rightarrow 0_+} MC(\varepsilon) > 0$. Then, equation (29) yields:

$$\lim_{\varepsilon \rightarrow 0_+} \int u(\theta, CE(\varepsilon)) dF(\theta) > qu_H(0) + (1-q)u_L(0) = \int u(\theta, 0) dF(\theta),$$

where the last equality follows from Bayes' rule. Since u is continuous and increasing in money, this implies that $\lim_{\varepsilon \rightarrow 0_+} CE(\varepsilon) > 0$. Hence, $\lim_{\varepsilon \rightarrow 0_+} \pi(\varepsilon)/\varepsilon = -\lim_{\varepsilon \rightarrow 0_+} CE(\varepsilon)/\varepsilon < 0$. ■

Proof of Proposition 15: Since $MC(\varepsilon) = 0$ for all ε , equation (17) becomes

$$\int u(\theta, CE(0)) dF(\theta) = qu_H(\varepsilon H) + (1-q)u_L(\varepsilon L) + z\chi(\varepsilon H, \varepsilon L). \quad (30)$$

Substituting $\chi(0, 0) = 0$, yields

$$\int u(\theta, CE(0)) dF(\theta) = qu_H(0) + (1-q)u_L(0).$$

Therefore, Bayes' rule implies that $\int u(\theta, CE(0)) dF(\theta) = \int u(\theta, 0) dF(\theta)$ and, because u is strictly increasing in money, $\pi(0) = -CE(0) = 0$.

Differentiating equation (30), it follows that

$$CE'(0) = \frac{qu'_H(0)H + (1-q)u'_L(0)L + z(H-L)[u'_H(0) - u'_L(0)]}{qu'_H(CE) + (1-q)u'_L(CE)}.$$

Substituting $qH + (1-q)L = 0$, yields

$$CE'(0) = K[u'_H(0) - u'_L(0)],$$

where $K = \frac{H}{qu'_H(0) + (1-q)u'_L(0)} \left(q + \frac{z}{1-q} \right) > 0$. Thus, applying L'Hospital, we obtain

$$\lim_{\varepsilon \rightarrow 0_+} \pi(\varepsilon)/\varepsilon = -CE'(0) = -K[u'_H(0) - u'_L(0)],$$

which concludes the proof. ■

D2 Remarks and Examples

Proof of the claim in Remark 1: Let $\hat{\mu}$ and μ denote the cumulative distribution functions of $\hat{\theta}_{\hat{\sigma}} \in \{\hat{\theta}_L, \hat{\theta}_{\emptyset}, \hat{\theta}_H\}$ and $\theta_{\sigma} \in \{\theta_L, \theta_H\}$, respectively. $\hat{\theta}_{\hat{\sigma}}$ second-order stochastically dominates θ_{σ} if, for any concave function $g : \Theta \rightarrow \mathbb{R}$,

$$\int g(\hat{\theta}_{\hat{\sigma}}) d\mu(\hat{\theta}_{\hat{\sigma}}) \geq \int g(\theta_{\sigma}) d\mu(\theta_{\sigma}). \quad (31)$$

But

$$\int g(\theta_{\sigma}) d\mu(\theta_{\sigma}) = qg(\theta_H) + (1-q)g(\theta_L), \text{ and}$$

$$\begin{aligned} \int g(\hat{\theta}_{\hat{\sigma}}) d\mu(\hat{\theta}_{\hat{\sigma}}) &= q(m_H + \eta_H)g(\theta_H) + [q(1 - m_H - \eta_H) + (1-q)(1 - m_L - \eta_L)]g(\hat{\theta}_{\emptyset}) \\ &\quad + (1-q)(\eta_L + m_L)g(\theta_L). \end{aligned}$$

Substituting in inequality (31) and dividing by $q(1 - m_H - \eta_H) + (1 - q)(1 - m_L - \eta_L)$, we obtain:

$$g(\alpha(m_L, m_H)\theta_H + [1 - \alpha(m_L, m_H)]\theta_L) \geq \alpha(m_L, m_H)g(\theta_H) + [1 - \alpha(m_L, m_H)]g(\theta_L),$$

which is true because g is concave. ■

Example 5: It is helpful to separate the analysis in 2 cases: (i) $q \geq \frac{2}{5}$, and (ii) $q < \frac{2}{5}$. In case (i), self 2 chooses a high ex-post action, $b(\emptyset) = b_H$ when she expects self 1 to manipulate her memory, $m_L = -\frac{2}{3}$. In case (ii), she chooses a low ex-post action, $b(\emptyset) = b_L$ when she expects $m_L = -\frac{2}{3}$.

Case (i):

The DM chooses to manipulate her memory if

$$\begin{aligned} & \left(1 - \frac{2}{3}\right) [v_L(b_L) + \tau(L)] \\ & + \frac{2}{3} \{ \alpha [v_H(b_L) + \tau(L)] + (1 - \alpha) [v_L(b_L) + \tau(L)] \} \\ & - \psi_L\left(-\frac{2}{3}\right) > v_L(b_L) + \tau(L), \end{aligned}$$

where α denotes the weight implied by Bayes' rule. This inequality is satisfied if and only if $\alpha > \frac{3}{32}$. Substituting the definition of α , we obtain $q > \frac{2}{31}$, which is satisfied since $q \geq \frac{2}{5} > \frac{2}{31}$.

The ex-ante expected utility from the signal is thus

$$\begin{aligned} & q [v_H(b_H) + \tau(H)] + \frac{2}{3} (1 - q) [v_L(b_H) + \tau(L)] \\ & + (1 - q) \left(1 - \frac{2}{3}\right) [v_L(b_L) + \tau(L)] - (1 - q) \psi_L\left(-\frac{2}{3}\right), \end{aligned}$$

which is equal to $\frac{83q+1}{12}$. If the DM makes an uninformed decision, she obtains an ex-ante utility of

$$\begin{aligned} & q [v_H(b_H) + \tau(H)] + (1 - q) [v_L(b_H) + \tau(L)] \text{ if } b = b_H, \text{ and} \\ & q [v_H(b_L) + \tau(H)] + (1 - q) [v_L(b_L) + \tau(L)] \text{ if } b = b_L. \end{aligned}$$

Thus, her utility is $7q$ if $q \geq \frac{1}{2}$, and $5q + 1$ if $q < \frac{1}{2}$. The surplus from observing the signal is then

$$\frac{83q+1}{12} - \max\{7q, 5q+1\} = \begin{cases} \frac{1-q}{12} & \text{if } q \geq \frac{1}{2}, \\ \frac{23q-11}{12} & \text{if } q < \frac{1}{2}, \end{cases},$$

which is positive if and only if $q > \frac{11}{23}$.

Case (ii): In this case, because $b(\emptyset) = b_L$, the signal has no value. Therefore, the DM is always (weakly) better off by not observing the signal. In particular, since she exerts memory manipulation if $q \geq \frac{2}{31}$, the surplus is strictly negative for $q > \frac{2}{31}$ and it is equal to zero if $q < \frac{2}{31}$.

Example 6: Following the same steps as in Example 5, it is straightforward to show that the surplus from observing the signal S is equal to

$$\begin{aligned} & (1 - q) \left(\frac{3}{4} - \bar{m}\right) \text{ if } q \geq \frac{1}{2}, \text{ and} \\ & 2q - 1 + (1 - q) \left(1 - \bar{m} - \frac{1}{4}\right) \text{ if } q < \frac{1}{2}. \end{aligned}$$

Therefore, S is always positive when $q \geq \frac{1}{2}$. Furthermore, for $q < \frac{1}{2}$, it is positive if and only if

$$2q - 1 + (1 - q) \left(1 - \bar{m} - \frac{1}{4} \right) \geq 0,$$

which simplifies to

$$\frac{3}{4} - \frac{1 - 2q}{1 - q} \geq \bar{m}.$$

Noting that $\bar{m}$ is distributed according to the c.d.f. Φ concludes the proof.

References

- Akerlof, G. A. and W. T. Dickens (1982). "The Economic Consequences of Cognitive Dissonance," *American Economic Review*, 72(3), 307-319.
- Allport, G. W. (1943). "The ego in contemporary psychology," *Psychological Review*, 50, 451-478.
- Anderson, J. R. (1995). *Learning and Memory: An Integrated Approach*, New York: John Wiley & Sons.
- Arkes, H. R., and C. Blumer (1985). "The psychology of sunk cost," *Organizational Behavior and Human Decision Processes*, 35, 124-140.
- Baron, R. A. (1999). "Counterfactual thinking and venture formation: The potential effects of thinking about what might have been," *Journal of Business Venturing* 15, 79-91.
- Bell, D. E. (1982). "Regret in decision making under uncertainty," *Operations Research*, 30, 961-981.
- Benabou, R. (2008). "Groupthink: Collective Delusions in Organizations and Markets," Mimeo., Princeton University.
- Benabou, R. and J. Tirole (2002). "Self-Confidence and Personal Motivation," *Quarterly Journal of Economics*, 117(3), 871-915.
- Benabou, R. and J. Tirole (2004). "Willpower and Personal Rules," *Journal of Political Economy*, 112 (4), 848-886.
- Benabou, R. and J. Tirole (2006a). "Belief in a Just World and Redistributive Politics," *Quarterly Journal of Economics*, 121(2), 699-746.
- Benabou, R. and J. Tirole (2006b). "Incentives and Prosocial Behavior," *American Economic Review*, 96(5), 1652-1678.
- Benabou, R. and J. Tirole (2006c). "Identity, Dignity and Taboos: Beliefs as Assets," Mimeo., Princeton University and Université de Toulouse.
- Berglas, S. and Baumeister, R. (1993). *Your Own Worst Enemy: Understanding the Paradox of Self-Defeating Behavior*, New York: BasicBooks.
- Bernheim, B. D. and R. Thomsen (2005). "Memory and Anticipation," *Economic Journal*, 115, 271-304.
- Billot, A, Chateauneuf, A., Gilboa, I., and Tallon, J.-M. (2000). "Sharing Beliefs: Between Agreeing and Disagreeing," *Econometrica*, 68(3), 685-694.
- Bodner, R. and D. Prelec. "Self-signaling in a neo-Calvinist model of everyday decision making," in *Psychology and Economics*, Vol. II. Brocas and J. Carillo (eds.), Oxford University Press, 2002.

- Brocas, I. and J. D. Carrillo (2008). "The Brain as a Hierarchical Organization," *American Economic Review*, forthcoming.
- Brockner, J. (1992). "The escalation of commitment to a failing course of action: toward theoretical progress," *Academy of Management Review*, 17, 39–61.
- Brockner, J., Houser, R., Birnbaum, G., Lloyd, K., Deitcher, J., Nathanson, S., and J. Z. Rubin (1986). "Escalation of Commitment to an Ineffective Course of Action: The Effect of Feedback Having Negative Implications for Self-Identity," *Administrative Science Quarterly*, 31, 109-126.
- Brunnermeier, M. K., and J. A. Parker (2005). "Optimal Expectations," *American Economic Review*, 95, 1092-1118.
- Budescu, D. V., Weinberg, S., and T. S. Wallsten (1988). "Decisions based on numerically and verbally expressed uncertainties," *Journal of Experimental Psychology: Human Perception and Performance*, 14, 281-294.
- Budescu, D. V., Kuhn, K. M., Kramer, K. M., and T. R. Johnson (2002). "Modeling certainty equivalents for imprecise gambles," *Organizational Behavior and Human Decision Processes*, 88, 748-768.
- Byrne, C. C. and Kurland, J. A. (2001). "Self-deception in an Evolutionary Game," *Journal of Theoretical Biology*, 212 (4), 457-480.
- Camerer, C. F. and R. A. Weber (1999). "The econometrics and behavioral economics of escalation of commitment: a re-examination of Staw and Hoang's NBA data," *Journal of Economic Behavior & Organization*, 39, 59-82.
- Caplin, A. and K. Eliaz (2003). "AIDS Policy and Psychology: A Mechanism-Design Approach," *RAND Journal of Economics*, 34(4), 631-646.
- Caplin, A. and J. Leahy (2001). "Psychological Expected Utility Theory and Anticipatory Feelings," *Quarterly Journal of Economics*, 116 (1), 55-79.
- Caplin, A. and J. Leahy (2004). "The Supply of Information by a Concerned Expert," *Economic Journal*, 114, 487-505.
- Compte, O., and and Andrew Postlewaite. (2004) "Confidence-Enhanced Performance," *American Economic Review*, 94, 1536-1557.
- Chau, A. W., & Phillips, J. G. (1995) "Effects of perceived control upon wagering and attributions in computer blackjack," *Journal of General Psychology*, 122, 253–269.
- Chow, C. C. and R.K. Sarin (2001). "Comparative Ignorance and the Ellsberg Paradox," *Journal of Risk and Uncertainty*, 22, 129-139.
- Dawson, E., Savitsky, K., and Dunning, D. (2006). "Don't Tell Me, I Don't Want To Know: Understanding People's Reluctance to Obtain Medical Diagnostic Information," *Journal of Applied Social Psychology*, 36, 751-768.
- Di Mauro, C. (2008). "Uncertainty Aversion vs. Competence: An Experimental Market Study," *Theory and Decision*, 64, 301-331.
- Dow, J. (1991). "Search Decisions with Limited Memory," *Review of Economic Studies*, 58, 1-14.
- Dunning, D. (1995). "Trait importance and modifiability as factors influencing self-assessment and self-enhancement motives," *Personality and Social Psychology Bulletin*, 21, 1297-1306.
- Eliaz, K. and R. Spiegler (2006). "Can anticipatory feelings explain anomalous choices of information sources?" *Games and Economic Behavior* 56, 87–104.

- Epstein, L. G. (1999). "A Definition of Uncertainty Aversion," *Review of Economic Studies*, 66, 579-608.
- Epstein, L. G. (2007). "Living with Risk," Rochester Center for Economic Research, Working Paper #534.
- Falk, A., Huffman, D., and U. Sunde (2006). "Self-Confidence and Search," IZA Discussion Paper #2525.
- Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.
- Fox, C. R., and A. Tversky, "Ambiguity Aversion and Comparative Ignorance," *Quarterly Journal of Economics*, 110, 585-603.
- Fox, C. R., and M. Weber. (2002). "Ambiguity aversion, comparative ignorance, and decision context," *Organizational Behavior and Human Decision Processes*, 88, 476-498.
- Fudenberg, D. and D. K. Levine (2006). "A Dual Self Model of Impulse Control," *American Economic Review*, 96, 1449-1476.
- Fudenberg, D. and D. K. Levine (2007). "Self Control, Risk Aversion, and the Allais Paradox," Working Paper, Harvard University and Washington University in St. Louis.
- Genesove, D. and C. Mayer (2001). "Loss Aversion And Seller Behavior: Evidence From The Housing Market," *Quarterly Journal of Economics*, 116, 1233-1260.
- Gollwitzer, P. M., Earle, W. B., and Stephan, W. G. (1982). "Affect as a determinant of egotism: Residual excitation and performance attributions," *Journal of Personality and Social Psychology*, 43, 702-709.
- Goodie, A. S. (2003). "The Effects of Control on Betting: Paradoxical Betting on Items of High Confidence With Low Value," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 598-610.
- Goodie, A. S. and D. L. Young (2007). "The skill element in decision making under uncertainty: Control or competence?," *Judgment and Decision Making*, 2, 189-203.
- Greenwald, A. G. (1980). "The Totalitarian Ego: Fabrication and Revision of Personal History," *American Psychologist* 35, 603-618.
- Grieco, D. and R. M. Hogarth (2004). "Excess entry, ambiguity seeking, and competence: An experimental investigation," Working Papers #778, Universitat Pompeu Fabra.
- Hanany, E. and P. Klibanoff (2007). "Updating preferences with multiple priors," *Theoretical Economics*, 2, 261-298.
- Harbaugh, R. (2008). "Prospect Theory or Skill Signaling?" Working Paper, Indiana University.
- Heath, C. and A. Tversky (1991). "Preference and belief: Ambiguity and competence in choice under uncertainty," *Journal of Risk and Uncertainty*, 4(1), 5-28.
- Hellman, M. E. and T. M. Cover (1973). "A Review of Recent Results on Learning with Finite Memory," *Problems of Control and Information Theory*, 221-227.
- Hilgard, E. R. (1949). "Human motives and the concept of self. *American Psychologist*," 4, 374-382.
- Hirshleifer, D. and I. Welch (2002). "An Economic Approach to the Psychology of Change: Amnesia, Inertia, and Impulsiveness," *Journal of Economics & Management Strategy*, 11, 379 - 421.
- Horswill, M. S. and F. P. McKenna (1999). "The effect of perceived control on risk taking," *Journal of Applied Social Psychology*, 29, 378-392.

- Hvide, H. (2002). "Pragmatic beliefs and overconfidence," *Journal of Economic Behavior and Organization*, 2002, 48, 15-28.
- Johnston, W. A. (1967). "Individual performance and self-evaluation in a simulated team," *Organizational Behavior and Human Performance*, 2, 309-328.
- Josephs, R. A., Larrick, R. P., Steele, C. M., and Nisbett, R. E. (1992). "Protecting the Self from the Negative Consequences of Risky Decisions," *Journal of Personality and Social Psychology*, 62, 26-37.
- Kahneman, D. and A. Tversky, 1979. "Prospect theory: An analysis of decision under risk," *Econometrica*, 47, 263-291.
- Kahneman, D., J. L. Knetsch., and R. H. Thaler (1990). "Experimental Tests of the Endowment Effect and the Coase Theorem," *Journal of Political Economy*, 98(6), 1325-1348.
- Keppe, H. and M. Weber (1995). "Judged Knowledge and Ambiguity Aversion," *Theory and Decision*, 39, 51-77.
- Kihlstrom, J. F., Beer, J. S., and S. B. Klein (2003). "Self and identity as memory," In: M. R. Leary and J. P. Tangney (Eds.), *Handbook of Self and Identity*. (Guilford Press: New York, NY).
- Kilka, M. and M. Weber (2000). "Home Bias in International Stock Return Expectation," *Journal of Psychology and Financial Markets* 1, 176-192.
- Korner, I. (1950) "Experimental Investigation of Some Aspects of the Problem of Repression: Repressive Forgetting." New York, NY: Contributions to Education, No. 970, Bureau of Publications, Teachers' College, Columbia University.
- Köszegi, B. (2003) "Health anxiety and patient behavior," *Journal of Health Economics*, 22, 1073-1084.
- Köszegi, B. (2006) "Ego Utility, Overconfidence, and Task Choice," *Journal of the European Economic Association*, 4, 673-707.
- Kuehberger, A. and Perner, J. (2003). "The role of competition and knowledge in the Ellsberg task," *Journal of Behavioural Decision Making*, 16, 181-191.
- Kunda, Z. and Sanitioso, R. (1989) "Motivated Changes in the Self-Concept," *Journal of Personality and Social Psychology*, 61, 884-897.
- Larrick, R. P. (1993) "Motivational Factors in Decision Theories: The Role of Self-Protection," *Psychological Bulletin*, 113, 440-450.
- Li, W. (2007). "Changing One's Mind when the Facts Change: Incentives of Experts and the Design of Reporting Protocols," *Review of Economic Studies*, 74, 1175-1194.
- List, J. (2003). "Does Market Experience Eliminate Market Anomalies?," *Quarterly Journal of Economics*, 118 (1), pages 41-71.
- List, J. and Haigh, M. S. (2005). "A simple test of expected utility theory using professional traders," *Proceedings of the National Academy of Sciences* 102, 945-948.
- Loomes, G. and Sugden, R. (1982). "Regret theory: An alternative theory of rational choice under uncertainty," *Economic Journal*, 92, 805-824.
- Lowenstein, G. A. (1987), "Anticipation and the valuation of delayed consumption," *Economic Journal*, 97, 666-684.
- Mather, M., E. Shafir, and M.K. Johnson (2003). "Memory and Remembering chosen and assigned options," *Memory & Cognition*, 31 (3), 422-433.
- Mill J. (1829). *Analysis of the Phenomena of the Human Mind*, (London: Baldwin and Craddock).

- Mukerji, S. (1998). "Ambiguity Aversion and Incompleteness of Contractual Form," *American Economic Review*, 88 (5), 1207-1231.
- Mullainathan, S. (2002). "A memory-based model of bounded rationality," *Quarterly Journal of Economics*, 117(3), 735-774.
- Philipson, T. and R. Posner (1995). "A Theoretical and Empirical Investigation of the Effects of Public Health Subsidies for STD Testing," *Quarterly Journal of Economics*, 110(2), 445-474.
- Piccione, M. and Rubinstein, A. (1997). "On the interpretation of decision problems with imperfect recall," *Games and Economic Behavior*, 20, 3-24.
- Prelec, D. (1998). "The probability weighting function," *Econometrica*, 66, 497-527.
- Prelec, D. (2008). "Self-delusion: A neuroeconomics model and fMRI evidence," Seminar on the Foundations of Human Social Behavior, University of Zurich.
- Quattrone G. A., and Tversky, A. (1984). "Causal Versus Diagnostic Contingencies: On Self-Deception and the Voter's Illusion." *Journal of Personality and Social Psychology*, 46, 237-248.
- Rabin, M. (1994). "Cognitive Dissonance and Social Change," *Journal of Economic Behavior and Organization*, 23, 177-194.
- Rabin, M. (2000). "Risk Aversion and Expected-Utility Theory: A Calibration Exercise," *Econometrica*, 68(5), 1281-1292.
- Rapaport, D. (1961). *Emotions and Memory*, (New York: Science Editions).
- Robbins, H. (1956). "A Sequential Decision Problem with a Finite Memory," *Proceedings of the National Academy of Sciences*, 42, 920-923.
- Safra, Z. and Segal, U. (2008). "Calibration Results for Non-Expected Utility Theories," *Econometrica*, forthcoming.
- Samuelson, P. A. (1963). "Risk and uncertainty: A fallacy of large numbers," *Scientia*, 6th Series 57, 1-6.
- Segal, U. and A. Spivak (1990). "First Order versus Second Order Risk Aversion," *Journal of Economic Theory*, 51, 111-125.
- Schelling, T. (1985). "The Mind as a Consuming Organ," In: J. Elster, ed. *The Multiple Self*. New York: Cambridge University Press.
- Sedikides, C., Green, J. D., and Pinter, B. T. (2004). "Self-protective memory," In: D. Beike, J. Lampinen, and D. Behrend, eds., *The self and memory*. Philadelphia: Psychology Press.
- Staw, B. M. (1976) "Knee Deep in the Big Muddy," *Organizational Behavior and Human Decision Process* 35, 124-140.
- Staw, B. M. (1981). "The Escalation of Commitment to a Course of Action," *The Academy of Management Review*, 6, 577-587.
- Steele, C. M. (1988). "The Psychology of Self-Affirmation: Sustaining the Integrity of the Self," In: L. Berkowitz (Ed.), *Advances in experimental social psychology*, (Vol. 21, 261-302). New York: Academic Press.
- Taylor, K. (1995) "Testing Credit and Blame Attributions as Explanation for Choices under Ambiguity," *Organizational Behavior and Human Decision Processes*, 64, 128-137.
- Taylor, S. E. and Gollwitzer, P. M. (1995). "The effects of mindset on positive illusions," *Journal of Personality and Social Psychology*, 69, 213-226.

- Thaler, R.H. (1980). "Toward a Positive Theory of Consumer Choice," *Journal of Economic Behaviour and Organization*, 1, 39-60.
- Thaler, R.H., and H.M. Shefrin (1981). "An Economic Theory of Self-control," *Journal of Political Economy*, 89, 392-406.
- Tirole, J. (2008). "Cognition and Incomplete Contracts," *American Economic Review*, forthcoming.
- Trivers, R. (2000). "The Elements of a Scientific Theory of Self-Deception," *Annals of the New York Academy of Sciences*, 907, 114-131.
- Tversky, A. and D. Kahneman (1992). "Advances in prospect theory: cumulative representation of uncertainty," *Journal of Risk and Uncertainty*, 5, 297-323.
- Van den Steen, E. (2004). "Rational Overoptimism (and Other Biases)," *American Economic Review*, 94, 1141-1151.
- Weinberg, B. A. (2006). "A Model of Overconfidence." Mimeo., Ohio State University.
- Wilson, A. (2003). "Bounded Memory and Biases in Information Processing," Mimeo., Princeton University.
- Zuckerman, M. (1979). "Attribution of success and failure revisited, or: The motivational bias is alive and well in attribution theory," *Journal of Personality*, 47, 245-287.