Imperfect Memory and Behavior under Risk

Daniel Gottlieb*

June 20, 2008

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 perceived characteristics and can, to some extend, manipulate their memories. It leads to a non-expected utility representation and provides a unified explanation for several empirical regularities, including non-linear probability weights, first-order risk aversion, the uncertainty effect, the endowment effect, and the sunk cost fallacy.

1 Introduction

Choices with uncertain outcomes are an important part of a person’s life. Most of the times, the outcomes are at least partially determined by the person’s characteristics. Therefore, they affect how one views himself and how he is viewed by others. Choices that turn out to be wrong typically lead to self-doubts while choices that turn out to be right enhance the person’s self-image.

A person that cares about self-image has an incentive to manipulate recollections and beliefs. Of course, this incentive does not matter if memory is perfect. This paper analyzes how the concern for self-image affects an individual’s behavior under risk when memory is imperfect. The individual is assumed to behave as a standard expected utility maximizer except for the fact that he has imperfect memory.

Memory management leads the decision-maker to avoid lotteries whose outcomes generate memory manipulation. Theorem 1 (Section 2) shows that preferences over signals $\Sigma$ whose outcomes $\sigma \in \{L, H\}$ are correlated with the person’s characteristics can be represented by a utility function of the form

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

where $u_\sigma$ is the expected utility given $\sigma$ and $q$ is the probability of observing $\sigma = H$. $w$ is a probability weighting function such that $w(q) \leq q$, with strict inequality whenever there is memory manipulation. Therefore, outcomes that lead to memory manipulation are worth less than predicted by the standard expected utility model. Subsection 2.5 employs this representation to discuss the demand for information.

Section 4 augments the model to allow for lotteries over money. Theorem 2 shows that preferences over lotteries $\mathcal{L}$ with monetary outcomes $t \in \{H, L\}$ can be represented by

$$U(\mathcal{L}) = w(q)v_H + [1 - w(q)]v_L,$$

1

Department of Economics, MIT. Preliminary version.
where \( v_t \) is the decision-maker’s utility in the state where \( t \) occurs. The probability weight \( w(q) \) differs from the probability \( q \) in two respects. First, \( w(q) \) is lower when outcomes lead to memory manipulation. Second, the larger the degree of complementarity between ability and money, the larger \( w(q) \) will be. These preferences provide an explanation for non-linear weighting functions based on memory manipulation.

As in other models that admit representations by non-linear weighting functions, the decision-maker may exhibit first-order risk aversion (Subsection 4.2). An agent with first-order risk aversion rejects small 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 endowment (endowment effect). However, differently from other non-expected utility models, the departure from linear weighting functions in our model is directly related with the decision-maker’s perceived characteristics. This is consistent with experimental evidence suggesting that departures from expected utility theory are associated with the lotteries being correlated with the decision-maker’s skill or knowledge [e.g. Heath and Tversky (1991), Fox and Tversky (1995), Goodie (2003), and Goodie and Young (2007)].

Subsection 4.3 presents a self-deception explanation for the endowment effect. The main idea is that successful trading often 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, he must also estimate the future prices of the goods. Thus, the outcome of the trade is informative about the person’s skills. Since decision-makers avoid information correlated with skills, they will only accept to trade if the expected benefit from trade is above a certain positive threshold.

In Subsection 4.4, it is shown that memory manipulation may cause people to value a lottery less than the worst possible monetary outcome from that lottery. Known as the uncertainty effect, this result was documented by Gneezy, List, and Wu (2006), who claim that it contradicts “virtually all models of risky choice”.

Subsection 4.5 provides a self-views rationale for the existence of sunk cost effects. According to this explanation, which is consistent with arguments made in the psychology literature, abandoning a project usually involves admitting that a wrong decision was made. Thus, revising one’s position is often informative about the decision-maker’s skills or knowledge. Sunk decisions may influence current behavior if the agent wants to avoid information about his characteristics.

Section 3 considers a repeated environment. The agent observes a sequence of signals that are informative about her characteristics and engages in memory manipulation after each realization. It is shown that the agent’s behavior and attitude towards risk converge to that implied by expected utility theory as the number of signals grows. This result is consistent with the argument that people do not exhibit ambiguity aversion over events that have been observed several times and that experts are subject to much less biases than beginners [e.g. List (2003)].

Sections 1.1 and 1.2 briefly review the psychological evidence on memory and the related literature in Economics.

1.1 An Overview

“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 committed to a task, he behaves quite differently. In the first condition his ego is not engaged; in the second, it is.”

\(^1\)See Subsection 4.1 for a more detailed discussion.
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 as being “smart”, “knowledgeable”, and “nice”. Information conflicting with this image is usually ignored or denied.

People are 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 characteristics which they believe to be correlated with success in professional or personal life [Kunda and Sanitioso (1989)]. Failure is usually attributed to external factors while success tends to be attributed to one’s own actions [Zuckerman (1979)].

Self-assessments and the memory are intrinsically connected. In his *Essay Concerning Human Understanding*, Locke (1690) identified the self with memory. Mill (1829) 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). The discomfort felt when one is presented with evidence that conflicts with his or her beliefs and the resulting effort to distort one’s beliefs or opinions was called cognitive dissonance.

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 they simply like to think that they have these attributes. Second, as argued by Compte and Postlewaite (2004), a person may benefit from having overconfident beliefs in situations where emotions affect performance. Third, since “the best liar is the one who believes his own lies,” there may be a signaling value. It may also play a role as a credible self-promotion or self-exaggeration device. Fourth, there may be a motivational value of belief manipulation. As argued by Benabou and Tirole (2002), confidence in one’s ability may help the person take more ambitious goals and persist in adverse situations.

In this paper, I abstract from the exact reason why people may value a positive self-image. The model developed here is based on the two basic premises discussed above. First, that individuals have preferences over their perceived abilities. Second, that they can (to some extent) affect what they will remember. Apart from these assumptions, individuals are assumed to behave as in standard economic models. Their preferences satisfy the standard expected

---

\(^2\)See Van den Steen (2004) for a model of rational agents with differing priors that generates these biases.

\(^3\)See Schelling’s (1985) theory of the mind as a consuming organ.

\(^4\)As argued by Trivers (2000), “[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 formally modelled by Byrne and Kurland (2001).
utility axioms. Furthermore, they follow Bayes’ rule and, therefore, are aware of their memory imperfection. The main focus of the paper is on how memory manipulations affect the person’s choice under risk.

As demonstrated by the opening quote from Allport, 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, Heath and Tversky (1991) argued that the consequences of bets include the credit and blame associated with the outcome as well as with monetary payoﬀs. Therefore, psychic costs resulting from the person’s self-evaluation may be an important component of behavior under risk.

In this paper, it is shown how the idea of self-deception can lead to a uniﬁed theory of choice under risk that is consistent with economic phenomena such as ambiguity averse/seeking behavior, ﬁrst-order risk aversion, the endowment effect, the sunk cost fallacy, and the uncertainty effect.

1.2 Related Literature

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. In this paper, I will follow the latter approach and consider the case of rational decision makers subject to imperfect recall.

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.

An important special case of imperfect memory are models of limited 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 a ﬁnite number of outcomes (ﬁnite memory). In a series of papers, Cover and Hellman characterized optimal solutions to some ﬁnite memory problems.

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 (2004) 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. Bernheim and Thomadsen (2005) showed that memory imperfections and anticipatory emotions may lead to a resolution of Newcomb’s Paradox and sustain cooperation in the Prisoners Dilemma.

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 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, the idea of anticipatory emotions has been studied by Lowenstein (1987), Caplin and Leahy (2001 and 2004), and Köszegi (2006).  

2 Basic Framework

Consider a decision-maker (DM) who has preferences over her characteristics $\theta$. The agent’s characteristics $\theta$ may be interpreted as her skills or knowledge as well as a parameter of anticipatory utility. Let $\Theta$ be a non-empty subset of $\mathbb{R}$ representing the possible values of $\theta$ and denote by $F(.)$ the agent’s prior distribution of $\theta$. Preferences satisfy the standard axioms of expected utility so that there exists a strictly increasing von Neumann-Morgenstern utility function $u : \Theta \rightarrow \mathbb{R}$ representing DM’s preferences over characteristics $\theta$.

The individual acts in two periods: 1 and 2. In period 1, a signal $\sigma$, which can be either high (H) or low (L), is observed. Denote the probability of observing a high signal by $q \in (0, 1)$. A high signal is assumed to be more favorable than $\sigma = L$ in the sense of first-order stochastic dominance:

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

with strict inequality for some value of $\theta$.

The informational structure is represented in Figure 1. Following Benabou and Tirole (2002, 2006a) and Rabin (1994), I assume that the individual can, at a cost, manipulate her recollections. The DM remembers a signal $i \in \{L, H\}$ with probability

$$\eta_i + r_i - f_i.$$

The parameter $\eta_i \in [0, 1]$ is the agent’s "natural" rate of remembering signal $i$. It determines the probability that DM recollects the signal if she does not employ any manipulation effort. However, the DM is able to depart from the natural rate of forgetting the signal by exerting efforts $r_i$ and $f_i$, where $r_i \in [0, 1 - \eta_i]$ denotes an effort to remember and $f_i \in [0, \eta_i + r_i]$ denotes an effort to forget the signal. Engaging in memory manipulation $r_i$ and $f_i$ leads to a cost of $\psi_r (r_i) \geq 0$ and $\psi_f (f_i) \geq 0$.

The costs of memory manipulation $\psi_r$ and $\psi_f$ can be related to psychic costs (stress from repression), 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 the shadow costs of memory in a limited information framework. Remembering one signal with probability above the natural rate $\eta_i$ requires the 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 $\psi_r$). Similarly, forgetting a signal with probability above the natural rate $1 - \eta_i$ requires the individual to focus on confronting evidence which, again, restricts the amount of attention available to other potentially useful information.

---

7 See also Philipson and Posner (1995), Caplin and Eliaz (2003), and Köszegi (2003) for specific applications.

8 $\Theta$ can be continuous or discrete, as long as it contains at least two elements (otherwise, $\theta$ cannot be random). Note that we have not assumed that the agent has a correct prior distribution over $\theta$. Therefore, agents are allowed to hold optimistic or pessimistic beliefs about their characteristics.

9 The natural rates of learning and forgetting were initially studied by the German psychologist Hermann Ebbinghaus.

10 In Hvide (2002) and Brunnermeier and Parker (2005), the costs of manipulation arise endogenously through suboptimal decision-making.
Assumption 1 The cost of memory manipulation $\psi_i(.)$ is strictly increasing, convex, twice-continuously differentiable, and satisfies $\psi_i(0) = 0, \ i \in \{r, f\}$.

Assumption 1 leads to a cost of memory manipulation as depicted in Figure 2. Because memory manipulation is costly, the DM will never simultaneously choose $r_i > 0$ and $f_i > 0$.\(^{11}\) Furthermore, I assume that there is some positive probability of forgetting a high signal if the agent does not spend effort in order to remember it: $\eta_H < 1$.\(^{12}\)

Throughout the paper, I will also consider two alternative assumptions about the cost of manipulation:

Assumption 2a $\psi'_i(0) = 0, \ i \in \{r, f\}$ and $\eta_L > 0$.

Assumption 2b $\psi''_i(0) > 0, \ i \in \{r, f\}$.

Assumption 2a implies that the cost of a small amount of memory manipulation is of order 2 and that the space of feasible $f_L$'s is non-trivial. 2b implies that the cost of a small amount of memory manipulation it is of order 1. As will be described in Subsection 2.2, Assumption 2a ensures that all equilibria have positive amounts of manipulation whereas there might be equilibria with no manipulation under 2b.

\(^{11}\)Lingering over positive feedback may also be pleasant for a certain amount of time. In that case, $\eta_H \in [0, 1]$ can be interpreted as the rate that maximizes utility from lingering over positive feedbacks. The results from the paper remain unchanged if the pleasure from lingering over good news is "not too high". Otherwise, the agent may be information-seeking.

\(^{12}\)If $\eta_H = 1$, then the model becomes trivial. Since the agent always recalls high signals, she will perfectly infer that $\sigma = L$ was observed if she recollects $\hat{\sigma} = \varnothing$. Therefore, she will never engage in memory manipulation.
Figure 2: Cost of Memory

Note that the assumption that $\sigma = H$ first-order stochastically dominates $\sigma = L$ implies that the agent will never spend effort in order to forget a high signal or remember a low signal (i.e., $r_L = f_H = 0$). Hence, with no loss of generality, we can restrict the set of feasible manipulation efforts to $r_H \in [0, 1 - \eta_H]$ and $f_L \in [0, \eta_L]$.

The model can be interpreted as a formalization of the neurophysiological argument put forth by Trivers (2000). He notes that it takes about 20 ms for a nervous signal to reach the brain, although it is only registered in consciousness after 500 ms. Furthermore, stimuli received up to 100 ms before the event reaches consciousness can affect the content of the experience. Trivers (2000) argues that “this is all time in the world, so to speak, for emendations, changes, deletions, and enhancements to occur.” Thus, the date-1 self in the model can be interpreted as the person’s unconscious process of information manipulation.\(^{13}\)

However, not all belief manipulation occurs between the moment where a signal is received and when it enters consciousness. By allocating attention and rehearsing, an individual is constantly involved in some sort of memory manipulation. Tirole (2002) provides an interesting example of self-deception. After receiving a hostile referee report, one typically tends to search for contradictory evidence or excuses: “The referee is either biased or incompetent.” Then, the person (consciously or not) tends to avoid negative cues later on, hides the report, and does not talk about the paper for a while.

The model can also be seen as a conflict between a “hot” or “impulsive” self and a “cold” self. The hot (date-1) self wants to minimize current losses from negative information and maximize the current gains from positive information. The cold (date-2) self, wants to circumvent the manipulations made by the other self in order to make a correct inference. The hot self exerts efforts $r_H$ and $f_L$ in order to manipulate the beliefs of the cold self. Then, the cold self applies Bayes’ rule in order to filter these manipulations.\(^{14}\)

\(^{13}\)This interpretation assumes that the agent’s unconscious process is rational in the sense of taking into account the benefits and costs of memory manipulation. Prelec (2008) shows experimental evidence suggesting that memory manipulation seems to respond positively to its expected benefits.

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

\(^{14}\)In the context of intertemporal choice, several papers have proposed dual self models [c.f. Thaler and Shefrin...
As the following examples show, the model described above encompasses other models of imperfect memory.

**Example 1 (The Forgetfulness Model)** If $\eta_L = 1$, $\eta_H = 0$ and $\psi_r(r) = +\infty$ for all $r > 0$, the informational structure reduces to the one in Benabou and Tirole (2002). As shown in Figure 3, it can be interpreted as a model of bad news or no news. If the agent receives bad news, she can exert an effort $f_L \in [0, 1]$ in order to forget them.

If we reinterpret the state $\emptyset$ as the recollection of a high signal, then the model becomes one where the agent is able to convince herself that a low signal was a high signal.\(^{15}\) Hence, memory manipulation would allow the DM to believe that she received a signal $\sigma = H$. 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 signal was actually a good one. In the other states (both when they acknowledge that the signal was bad news or when they observe a good signal), no such activity is detected. Hence, Example 1 can be interpreted as the agent incurring in psychological costs when she tries to convince herself that a bad signal was actually a good one.

**Example 2 (The Limited Memory Model)** Let $\eta_L = \eta_H = 0$ so that the DM forgets any signal 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 order to store information. By spending a memory cost $\psi_r(r_i)$, she remembers the signal with probability $r_i$. Since, in (1981), Fudenberg and Levine (2006), and Brocas and Carrillo (2008)].

\(^{15}\)In this model, the agent would never choose to believe that a high signal was actually a low signal.
our model, the agent would never choose to remember a low signal, there is no loss of generality in assuming that effort can only be spent after a high signal.\footnote{The main difference with respect to Dow (1991) and Wilson (2004) is that I allow the DM to remember a signal with any probability \( r_L, r_H \in [0, 1] \). Hence, a higher \( r_i \) is related to having higher memory resources used to store the information. In their models, however, the agent either remembers or forgets the information for sure.}

The following example consists of a model where memory is exogenously imperfect.

**Example 3 (Exogenous Memory Model)** Let \( \psi_r (r_i) = +\infty \) for all \( r_i > 0 \) and \( \psi_f (f_i) = +\infty \) for all \( f_i > 0 \) so that the agent does not engage in endogenous memory manipulation: \( f_i^* = r_i^* = 0, i = L, H \). Let \( \eta_i < 1 \) so that the agent forgets signals with (exogenous) probabilities \( 1 - \eta_i > 0 \). If \( \eta_H > \eta_L \), memory is selective in the sense that good news are more likely to be remembered than bad news.

2.1 Multiself Game

I follow 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. The perfect Bayesian equilibrium (PBE) of this game between selves corresponds to the decision made by an agent with imperfect recall.\footnote{For the games considered here, the set of sequential equilibria coincides with the set of PBE. Piccione and Rubinstein (1997) also propose a “modified multiverse consistency” condition which, for the games considered here, leads to the same equilibria as the ones obtained by applying the multiverse consistency approach.}
restricted by the assumption that recollections are interpreted according to Bayes’ rule. Thus, the agent makes correct inferences about her characteristics given the recollections.

Equivalently, we can conceptualize a ‘second-period self’ that tries to make a correct inference about the agent’s characteristics. The period-2 self chooses beliefs so as to minimize a quadratic loss function:

$$u^* = \arg \min_{\beta} \int -\beta^2 R(\beta) f(\beta|\sigma) d\beta;$$

The solution to this program is $u^* = \int u(\theta) f(\theta|\sigma) d\theta$, which is the Bayes estimator of $u(\theta)$ given the recollection $\hat{\sigma}$. Thus, by minimizing the quadratic loss function, the period-2 self constrains the decision-maker to be Bayesian given her memory imperfection.

The extensive form of the multiself game is represented in Figure 5. In period 0, nature plays $\sigma = H$ with probability $q$ and $\sigma = L$ with probability $1 - q$. Then, the date-1 self decides the amount of memory manipulation. Given $\sigma = L$ and manipulation effort $f_L$, nature plays $\hat{\sigma} = L$ with probability $\eta_L - f_L$ and $\hat{\sigma} = \emptyset$ with probability $1 - \eta_L + f_L$. Similarly, given $\sigma = H$ and $r_H$, nature plays $\hat{\sigma} = H$ with probability $\eta_H + r_H$ and $\hat{\sigma} = \emptyset$ with the complementary probability. Then, the date-2 self makes an inference about the characteristics $\theta$ so as to minimize a quadratic loss function given the recollections $\hat{\sigma}$.

An important assumption for the multiself approach is that DM cannot commit to an ex-
ante strategy. In the present model, memory manipulation leads to externalities among different selves. Forgetting a low signal increases the payoff of the agent after the signal was realized. However, this strategy would have left the agent worse off had a high signal occurred since, in this case, the expected payoff when a signal is forgotten is lower. The multiself approach assumes that, given a low realization, the self does not take into account the effect of memory manipulation on her utility had she observed a high signal instead.

**Definition 1** A perfect Bayesian equilibrium (PBE) of the game is a strategy profile \( r_H^*, f_L^* \) and posterior beliefs \( F(\cdot|\hat{\sigma}) \) such that:

1. \( f_L^* \in \arg \max_{f_L} (\eta_L - f_L) \int u(\theta) dF(\theta|\hat{\sigma} = L) + (1 - \eta_L + f_L) \int u(\theta) dF(\theta|\hat{\sigma} = \emptyset) - \psi_f(f_L) \)

2. \( r_H^* \in \arg \max_{r_H} (\eta_H + r_H) \int u(\theta) dF(\theta|\hat{\sigma} = H) + (1 - \eta_H - r_H) \int u(\theta) dF(\theta|\hat{\sigma} = \emptyset) - \psi_r(r_H) \)

3. \( \forall x \in \{L, H, \emptyset\}, F(\theta|\hat{\sigma} = x) \) is obtained by Bayes’ rule if \( \Pr(\hat{\sigma} = x|f_L^*, r_H^*) > 0 \).

1 and 2 are the standard perfection conditions. Condition 3 implies that beliefs are given by Bayes’ rule. Denote the expected utilities given \( \sigma = H \) and \( \sigma = L \) by

\[
u_H \equiv \int u(\theta) dF(\theta|\sigma = H), \quad \text{and} \quad u_L \equiv \int u(\theta) dF(\theta|\sigma = L).
\]

Given the recollection \( \hat{\sigma} = H \), the date-2 self infers that a high signal was observed in date 1. Hence, the expected utility of the date-1 self conditional on \( \hat{\sigma} = H \) is \( u_H \). Similarly, the expected utility of the date-1 self conditional on \( \hat{\sigma} = L \) is \( u_L \). If the DM forgets which signal was observed at date 1 (i.e., she recollects \( \hat{\sigma} = \emptyset \)), then there is a probability \( (1 - q)(1 - \eta_L + f_L) \) that \( \sigma = L \) was observed and a probability \( q(1 - \eta_H - r_H) \) that \( \sigma = H \) was observed. Thus, the expected utility given \( \hat{\sigma} = \emptyset \) is

\[
u_\emptyset \equiv \alpha(f_L, r_H) \times u_H + [1 - \alpha(f_L, r_H)] \times u_L, \tag{1}
\]

where \( \alpha(f_L, r_H) \equiv \frac{q(1 - \eta_H - r_H)}{q(1 - \eta_H - r_H) + (1 - q)(1 - \eta_L + f_L)} \) is the conditional probability of \( \sigma = H \) implied by Bayes’ rule.

**Remark 1** Denote the expected ability conditional on the observed signal \( \sigma \) by \( \theta_\sigma \) and the expected ability conditional on the recollections \( \hat{\sigma} \) by \( \theta_\hat{\sigma} \). \( \theta_\hat{\sigma} \) is "less variable" than \( \theta_\sigma \) in the sense of second-order stochastic dominance (see Appendix A). Therefore, since \( \theta_\sigma \) is the Bayes estimate of \( \theta \) given the signal \( \sigma \), forgetfulness implies that the decision-maker updates observed signals \( \sigma \) less than implied by Bayes’ rule. This result is consistent with experimental evidence from Falk, Huffman, and Sunde (2006).

### 2.2 Memory manipulation

After observing a low signal, the date-1 self has expected utility \((\eta_L - f_L) u_L + (1 - \eta_L + f_L) u_\emptyset - \psi_f(f_L)\). Let \( f_L^* \) and \( r_H^* \) denote the amount of memory manipulation that the date-2 self believes that was employed in time 1. Note that \( f_L^* \) and \( r_H^* \) are taken as given by the date-1 self when choosing the amount of memory manipulation to exert.

Using equation (1), the expected utility after a low signal can be written as

\[
u_L + \alpha(f_L^*, r_H^*) (1 - \eta_L + f_L) \Delta u - \psi_f(f_L), \tag{2}
\]
where $\Delta u \equiv u_H - u_L > 0$ denotes the utility gain from observing a high signal instead of a low signal. The period-1 self chooses the amount of manipulation effort $f_L$ so as to maximize the expression in (2).

Analogously, the expected utility of the date-1 self after observing a high signal is $(\eta_H + r_H) u_H + (1 - \eta_H - r_H) u_0 - \psi_r (r_H)$. Substituting equation (1), we obtain

$$u_L + \{(\eta_H + r_H) [1 - \alpha (f_L^*, r_H^*)] + \alpha (f_L^*, r_H^*)\} \Delta u - \psi_r (r_H).$$

(3)

The date-1 self chooses the amount of effort $r_H$ that maximizes the expression in (3).

Condition 3 for a PBE implies that the period-2 self correctly infers the amount of self-deception exerted in period 0 so that $f_L^* = f_L$ and $r_H^* = r_H$. The following proposition establishes the existence of a PBE and considers the possibility of equilibria at the boundary.

**Proposition 1** There exists a PBE. Suppose that $\psi_i$ is strictly convex, $i \in \{L, H\}$. Then, in any PBE:

1. If Assumption 2a holds, then $f_L^* > 0$, and $r_H^* > 0$,

2. If Assumption 2b holds, then

$$\psi'_f (0) \geq \Delta u \implies f_L^* = 0, \text{ and}$$

$$\psi'_r (0) \geq \Delta u \implies r_H^* = 0.$$

Memory manipulation is ex-post desirable because the DM increases the probability of either remembering a high signal or forgetting a low signal, which increases her payoff. The incentives to engage in memory manipulation are higher the higher the payoff gain from a high signal compared to a low signal, $\Delta u$. The following lemma shows that whenever $\psi'_r (0) > 0$ and $\psi'_f (0) > 0$ we can split the range of $\Delta u$ in three intervals. When $\Delta u$ is sufficiently low, there is a unique equilibrium where the DM does not engage in memory manipulation. For intermediate values of $\Delta u$, there exist both equilibria where she engages in memory manipulation and equilibria where she does not. And, when $\Delta u$ is high enough, only equilibria where the DM exerts memory manipulation exist.

**Lemma 1** Suppose Assumption 2b holds. There exist $\lambda$ and $\tilde{\lambda}$ with $\tilde{\lambda} \geq \lambda > 0$ such that in any PBE,

$$\Delta u < \lambda \implies \max \{f_L^*, r_H^*\} = 0, \text{ and}$$

$$\Delta u > \tilde{\lambda} \implies \max \{f_L^*, r_H^*\} > 0.$$

Furthermore, when $\lambda \leq \Delta u \leq \tilde{\lambda}$, there is a PBE where $\max \{f_L^*, r_H^*\} > 0$ and a PBE where $\max \{f_L^*, r_H^*\} = 0$.

Next, we characterize the PBE in the models of Examples 1 and 2.

### 2.2.1 The forgetfulness model

Consider the forgetfulness model of Example 1. Because $r_H^* = 0$ we redefine the Bayesian weighting function as $\alpha (f_L) \equiv \frac{q}{q+(1-q)f_L}$. Given a low signal in date 1, the agent solves

$$\max_{f_L} (1 - f_L) u_L + f_L \left\{ \alpha (f_L) \times u_H + [1 - \alpha (f_L)] \times u_L \right\} - \psi_f (f_L).$$

(4)

12
Applying Kuhn-Tucker’s theorem, and substituting the equilibrium condition $f_L = f^*_L$, we obtain
\[
\left[\frac{q}{q + (1 - q) f^*_L}\right] \Delta u = \psi'_f (f^*_L),
\] (5)
in any interior equilibrium.

Let $f^*_L$ be implicitly defined by equation 5. From the implicit function theorem, such $f^*_L \in \mathbb{R}$ exists and is unique (see Figure 6).

The following proposition characterizes the PBE and presents some comparative statics results:

**Proposition 2** In the forgetfulness model, there exists a unique PBE given by equation (5) if
\[
\psi'_f (0) < \Delta u < \frac{\psi'_f (1)}{q},
\]
and
\[
f^*_L = 0 \text{ if } \Delta u \leq \psi'_f (0), \text{ and }
\]
\[
f^*_L = 1 \text{ if } \Delta u \geq \frac{\psi'_f (1)}{q}.
\]

Furthermore, the equilibrium amount of belief manipulation $f^*_L$ is:

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

The results above follow 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. Moreover, 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 the signal.
Hence, higher $q$'s increase the marginal benefit of self-deception which lead to an increase in manipulation $f_L.\textsuperscript{18}

2.2.2 The limited memory model

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

$$\max_{r_H} r_H u_H + \left(1 - r_H\right) \left\{ \alpha (r_H^*) u_H + \left[ 1 - \alpha (r_H^*) \right] u_L \right\} - \psi_r (r_H),$$

where, with some abuse of notation, I have defined the Bayesian weighting function as $\alpha (r_H) \equiv \frac{q(1-r_H)}{q(1-r_H)+1-q}$. In an interior equilibrium, we must have

$$\frac{1-q}{q(1-r_H^*)+1-q} \Delta u = \psi'_r (r_H). \tag{6}$$

Proceeding as in 2.2.1, we obtain:

**Proposition 3** *In the limited memory model, the PBE are characterized by equation (6) if*

$$\psi'_r (0) \leq \frac{1-q}{q(1-r_H^*)+1-q} \Delta u \leq \psi'_r (1),$$

$$r_H^* = 0 \text{ if } \Delta u \leq \frac{\psi'_r (0)}{1-q}, \text{ and}$$

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

An interesting feature of the limited memory model is the possibility of multiple equilibria. Since both sides of Equation (6) are increasing in $r_H^*$, there may be multiple interior equilibria. It may also simultaneously feature interior equilibria and corner equilibria or equilibria at both corners $r_H = 0$ and $r_H = 1.\textsuperscript{19}$

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 $\sigma = \sigma$. Therefore, she will have much more incentives to remember good signals. As we 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 the date-1 self believes that she will have engaged in more memory manipulation.\textsuperscript{20}

2.3 Ex-ante utility

In this subsection, we compute the decision-maker’s expected utility from observing the signal. In equilibrium, when the DM forgets the signal ($\hat{\sigma} = \emptyset$), she knows that there is a probability

\textsuperscript{18}Consistently with Claim 1 from Proposition 2, Prelec (2008) presents experimental evidence suggesting that self-deception is increasing in the benefits of manipulation.

\textsuperscript{19}For example, if $\psi'_r (1) \leq \Delta u \leq \frac{\psi'_r (0)}{1-q}$ there exist both an equilibrium with $r_H^* = 1$ and one with $r_H^* = 0$.

\textsuperscript{20}See Benabou and Tirole (2002) for a discussion of self-traps. The existence of multiple equilibria is interesting since there seems to be a large heterogeneity in the amount of self-deception accross different people [Prelec (2008)].
\( \alpha (f^*_L, r^*_H) \) that there was a high signal and \( 1 - \alpha (f^*_L, r^*_H) \) that there was a low signal. Therefore, Bayesian updating implies that the agent’s posterior beliefs take into account the effect of memory manipulation. On average, the only effects of engaging in self-deception are the manipulation costs \( \psi_f (f^*_L) \) and \( \psi_r (r^*_H) \). Of course, there is still an ex-post incentive to manipulate beliefs after she observes the signal. Nevertheless, the inability to commit not to engage in self-deception leads to a loss in (ex-ante) expected utility.

The decision maker’s ex-ante expected utility from observing the signal is the expected utility from characteristics minus the expected costs of memory manipulation:

\[
q u_H + (1-q) u_L - q \psi_r (r^*_H) - (1-q) \psi_f (f^*_L). \tag{7}
\]

The following result shows that, because the agent cannot commit not to manipulate her memory after the signal, she is better off by not observing it:

**Proposition 4** Let \( U^s \) denote the expected utility of observing the signal and \( U^{ns} \) the expected utility of not observing the signal. Then, \( U^{ns} \geq U^s \). Furthermore,

- under Assumption 2a, \( U^{ns} > U^s \);
- under Assumption 2b, there exist \( \lambda \) and \( \hat{\lambda} \) with \( \hat{\lambda} \geq \lambda > 0 \) such that in any PBE,

\[
\Delta u < \lambda \implies U^{ns} = U^s,
\]

\[
\Delta u > \hat{\lambda} \implies U^{ns} > U^s,
\]

and there is a PBE where \( U^{ns} > U^s \) and a PBE where \( U^{ns} = U^s \) when \( \lambda \leq \Delta u \leq \hat{\lambda} \).

In order to observe the signal, the individual requires a "participation premium" of \( q \psi_r (r^*_H) + (1-q) \psi_f (f^*_L) \).

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 choose not to engage in memory manipulation \( (r = f = 0) \). However, after receiving either a good or a bad signal, she would like to revise her previous choice and engage in manipulation. Hence, the ex-ante optimal strategy is time-inconsistent. This contrasts with decision problems with perfect recall, where ex-ante optimal strategies are time-consistent.\(^{21}\)

### 2.4 Probability weights

In this subsection, I will show that the model leads to a non-expected utility representation, where the decision-maker’s expected utility from observing the signal is equal to a weighted average of the utility in each state of the world \( \sigma \in \{L, H\} \). Hence, there exist weights \( w(q) \) such that the utility from observing a signal correlated with the DM’s characteristics is

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

where \( \Sigma \) is the lottery that has outcome \( \sigma = H \) with probability \( q \) and \( \sigma = L \) with probability \( (1-q) \). Clearly, the decision maker is an expected utility maximizer if \( w(q) = q \). If \( w(q) < q \), the agent is ambiguity averse.\(^{21}\)

\(^{21}\)See 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.
Simple algebraic manipulations of equation (7) yield
\[ w(q) = q - \frac{q \psi_r (r_H^*) + (1 - q) \psi_f (f_L^*)}{\Delta u}, \] (8)

Therefore, the model admits a non-expected utility representation, where the probability weights are given by equation (8).

**Theorem 1** The agent’s expected utility of observing the signal \( \sigma \in \{L, H\} \) can be represented by
\[ U(\Sigma) = w(q) u_H + [1 - w(q)] u_L, \] (9)

where \( u_\sigma \equiv \int u(\theta) dF(\theta|\sigma) \) and \( w \) is given by equation 8. Furthermore,

i. \( w(0) = 0 \),

ii. \( w(1) = 1 \),

iii. \( w(q) \leq q \) for all \( q \in [0, 1] \), and

iv. If \( r_H^* = 0 \) or \( f_L^* = 0 \), then \( w(q) \in [0, 1] \) \( \forall q \in [0, 1] \).

Theorem 1 implies that the weights \( w(q) \) are between 0 and 1 in Examples 1 and 2 and, therefore, can be thought of as probabilities (i.e., we can write \( U(\Sigma) = E_w [u(\theta)] \), where \( w \) is a probability measure). The expected utility representation is a special case of our model when there is no memory manipulation \( (r_H^* = f_L^* = 0 \text{ for all } q) \).

Whenever \( r_H^* > 0 \) or \( f_L^* > 0 \), the individual attributes a lower weight to the good state \( w(q) < q \). Hence, our model predicts ambiguity aversion whenever the individual engages in self-deception.

In order to compare lotteries with different probabilities of obtaining a high signal, it is important to keep the agent’s prior distribution over characteristics fixed. More specifically, when comparing two lotteries with different probabilities \( q \) of observing \( \sigma = H \), we want the choice to be driven by the different informational content of each lottery while keeping the DM’s prior distribution over characteristics \( f(\theta) \) fixed. In general, this implies that \( \Delta u \) is itself a function of \( q \). Therefore, in this section, we write \( \Delta u(q) \) in order to stress this dependence.22

Following Dow and Werlang (1992), define the degree of ambiguity aversion by
\[ c(q; u) \equiv 1 - w(q) - w(1 - q) \]
\[ = \frac{q \psi_r (r_H^*(q)) + (1 - q) \psi_f (f_L^*(q))}{\Delta u(q)} \]
\[ + \frac{(1 - q) \psi_r (r_H^*(1 - q)) + q \psi_f (f_L^*(1 - q))}{\Delta u(1 - q)}. \] (10)

Our model predicts that the most ambiguous events are those that lead to greatest amount of self-deception effort. The following proposition relates the amount of ambiguity aversion with the net marginal benefit (\( \Delta u \)) and the marginal cost of self-deception:

---

22See Appendix B for an example.
Proposition 5 Under Assumption 2a, $c(q;u) > 0$. Under Assumption 2b, there exist $\lambda$ and $\hat{\lambda}$ with $\hat{\lambda} \geq \lambda > 0$ such that, in any PBE,

$$
c(q;u) > 0 \text{ if } \max \{ \Delta u(q) ; \Delta u(1-q) \} > \hat{\lambda},$
$$
c(q;u) = 0 \text{ if } \max \{ \Delta u(q) ; \Delta u(1-q) \} < \lambda.
$$

Furthermore, when $\lambda \leq \max \{ \Delta u(q) ; \Delta u(1-q) \} \leq \hat{\lambda}$, there is a PBE where $c(q;u) > 0$ and a PBE where $c(q;u) = 0$.

Proof. The first part follows from Proposition 1 whereas the second part follows from Lemma 1.

2.5 Information Acquisition

In this subsection, I analyze the effects of memory manipulation for information acquisition when individuals have preferences over their characteristics (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. As before, let the space of possible characteristics $\Theta$ be a non-empty subset of $\mathbb{R}$ and let $F(.)$ denote the agent’s prior distribution of $\theta$. The DM has preferences that are represented by a strictly increasing von Neumann-Morgenstern utility function $u: \Theta \rightarrow \mathbb{R}$.

If the individual does not observe the signal, her utility is $\int u(\theta) dF(\theta)$. If she observes a signal $\sigma$, the utility conditional on $\sigma$ 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 $\sigma$. 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. In other words, in this standard model of ego utility, the fact that an individual has preferences over her expected skills does not influence her decision of acquiring information.

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 studies of information acquisition have, thus, assumed that utility functions were non-linear in probabilities [for 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), all these papers depart from the standard expected utility model by adopting the psychological expected utility model (PEU) of Caplin and Leahy (2001).\footnote{Philipson and Posner do not provide a justification for the assumption of a utility function that is non-linear in probabilities.}

As we have seen in the previous subsection, our model also leads to a utility function which is non-linear in probabilities. However, in our case, the non-linearity arises endogenously through memory manipulation. Therefore, our model can be seen as providing a cognitive foundation for a model of information acquisition.
More precisely, suppose the DM has the choice of collecting information about her characteristics before performing a task, which gives a random payoff. Knowing one’s skills lead to a more favorable distribution over payoffs. When will the decision maker prefer not to collect information about her characteristics?

As shown in Proposition 4, the DM requires a premium of \( q \psi_r (r^*_H) + (1 - q) \times \psi_f (f^*_L) \) in order to observe a signal. Hence, the individual is only willing to observe the signal if the expected costs of making an uninformed decision are greater than this participation premium. Otherwise, the agent prefers to obtain a lower expected payoff and avoid the costs of self-deception. This intuition will play a key role in the applications presented in Subsections (4.3), (4.4), and (4.5).

An immediate consequence of avoiding information correlated with one’s skills is the emergence 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 above predicts that a person may engage in such strategies if the expected costs are not too high.

3 Dynamic Model

Consider an infinitely repeated version of the game described previously. In each period \( n \in \{1, 2, 3, ..., N\} \), an independent draw of the signal \( \sigma_n \in \{H, L\} \) is made. Each \( \sigma_n \) has distribution given by \( \Pr (\sigma_n = H | \theta) \) and \( \Pr (\sigma_n = L | \theta) \), where \( \theta \) is the agent’s ‘true’ characteristics. The parameter \( \theta \) is not known. Instead, the decision-maker has a prior \( F(\theta) \) about its distribution. Hence, the prior over the distribution of the signal \( \sigma_n \) is

\[
\Pr (\sigma_n = i) = \int \Pr (\sigma_n | \theta = i) dF(\theta),
\]

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

After observing \( \sigma_n \in \{H, L\} \), the decision-maker engages in memory manipulation. She recollects a signal \( \hat{\sigma}_n \in \{\emptyset, L, H\} \). As in the static game, this is modeled through a different self acting each time an information is forgotten.

A history at time \( n \) is a sequence of recollections \( h^{n-1} = (\hat{\sigma}_1, \hat{\sigma}_2, ..., \hat{\sigma}_{n-1}) \in \{\emptyset, L, H\}^{n-1} \). In each period, the stage-1 self is a short-run player that chooses actions \( (r_H, f_L) : \{\emptyset, L, H\}^{n-1} \rightarrow [0, 1] \times [0, 1] \) to maximize the discounted sum of all future stage-game payoffs. The discount rate is \( \delta \in (0, 1) \). The stage-2 self always applies Bayes’ rule. Denote the sequence of (history-dependent) manipulations by \( r \equiv \{r_H (h^n) ; \forall h^n, n \} \) and \( f \equiv \{f_L (h^n) ; \forall h^n, n \} \).

Definition 2 A perfect Bayesian equilibrium (PBE) of the repeated game is a strategy profile \((r, f)\) and posterior beliefs \(F(\cdot | \hat{\sigma})\) such that for all \( h^n \) and all \( n = 1, 2, ..., N\):

1. \( f_L (L, h^{n-1}) = \arg \max_{f_L} (\eta_L - f_L) [\int u(\theta) dF(\theta | L, h^{n-1}) + \delta V(L, h^{n-1})] + (1 - \eta_L + f_L) [\int u(\theta) dF(\theta | \emptyset, h^{n-1}) + \delta V(\emptyset, h^{n-1})] - \psi_f (f_L) \)
2. \( r_H (H, h^{n-1}) = \arg \max_{r_H} (\eta_H + r_H) [\int u(\theta) dF(\theta | H, h^{n-1}) + \delta V(H, h^{n-1})] + (1 - \eta_H - r_H) [\int u(\theta) dF(\theta | \emptyset, h^{n-1}) + \delta V(\emptyset, h^{n-1})] - \psi_r (r_H) \)
3. \( F(\theta | h^n) \) is obtained by Bayes’ rule if \( \Pr (h^n | f, r) > 0 \).

\(^{24}\)For simplicity, assume that, if the agent chooses not to collect information, the payoffs are uninformative about her skills.
4. The continuation payoff $V$ satisfies \(^{25}\)

$$V (h^n) = \frac{1 - \delta^{N-n}}{1 - \delta} \times \int u (\theta) \, dF (\theta| h^n) - \sum_{s=0}^{N-n} \delta^{s} E_{h^{n+s}|h^n} \left[ \Pr (\sigma_{t+s} = H| h^n) \times \psi_r (r_H (h^{n+s})) + \Pr (\sigma_{t+s} = L| h^n) \times \psi_f (f_L (h^{n+s})) \right].$$

We are interested in the PBE of the game when $N$ is large. Let $\hat{u}_n (h^n)$ denote the Bayes estimator of $u (\theta)$ given history $h^n$.

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

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

I assume that $\eta_H > 0$ and that there exists some $\bar{f} < \eta_L$ with $\psi_F (\bar{f}) \geq \sup_{\theta} \{u (\theta)\} - \inf_{\theta} \{u (\theta)\}$. These assumptions ensure that the DM never forgets a signal $\sigma_n \in \{L, H\}$ with probability 1. \(^{26}\) The first issue is whether the Bayes estimator of $u (\theta)$ is consistent. In other words, does the DM eventually learn her true type after observing a sufficiently large number of signals?

If memory manipulation were exogenous (or, at least, constant), the answer would be immediate. Because, in this case, the recollections would be i.i.d., Doob’s Consistency theorem would imply that $\hat{u}_n (h^n)$ converges to $u (\theta)$. This is formally stated in the following lemma:

**Lemma 2** Suppose $r_n (h^{n-1}) = \bar{r}$ and $f_n (h^{n-1}) = \bar{f}$ for all $h^{n-1}, n$ and let $N \to \infty$. Then $\hat{u}_n \to u (\theta)$ 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 $\sigma_n$ are i.i.d., memory manipulation leads to non-independent and non-identically distributed recollected signals $\hat{\sigma}_n$. 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.

In order to show that this intuition is correct, I will use the following result:

**Lemma 3** For any fixed history $h^n$, $F (\theta| h^n; r, f)$ is increasing in $r$ and in $f$.

The Lemma above implies that, conditional on reaching each history, the agent always prefers that she had not engaged in memory manipulation. Because the agent is ultimately concerned about $\sigma_n$, $F (\theta| h^n; r, f)$ is not a function of $r$ and $f$ 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 if she had not manipulated memory (since memory manipulation increases the relative probability of arriving at $\hat{\sigma}_n = \emptyset$ after a low signal $\sigma_n = L$). Hence, $F (\theta| h^n; 0, 0)$ first-order stochastically dominates $F (\theta| h^n; r, f)$ for all $r, f$. \(^{27}\)

\(^{25}\)From the law of iterated expectations, $E \left[ E \left[ u (\theta) \mid h^{n+s} \right] \mid h^n \right] = E \left[ u (\theta) \mid h^n \right]$ for all histories $h^{n+s}$ that follow history $h^n$.

\(^{26}\)Either one of these conditions are needed to ensure identification. If $\eta_H = 0$ and $f_L (h^t) = \eta_L$, then $r_H (h^t) = 0$ for all $h^t$ implies that $\hat{\sigma} = \emptyset$ so that the bayesian posterior is equal to the prior. Therefore, there is no hope for the Bayes estimator to be consistent.

This assumption is not satisfied in the model of Example 2.2.1 ($\eta_H > 0$ is violated). However, it is straightforward to adjust the arguments from this section to establish the same results for that model.

\(^{27}\)The first-order dominance (FOSD) is for fixed $h^t$. Since the probability of each history is itself a function of $r$ and $f$, it does not follow that there is unconditional FOSD.
A straightforward implication of Lemma 3 is that:

\[ E\left[u(\theta) \mid h^n; 1 - \eta_H, \tilde{f}\right] \leq E\left[u(\theta) \mid h^n; r, f\right] \leq E\left[u(\theta) \mid h^n; 0, 0\right], \quad (11) \]

for all \( r \) and \( f \) and all histories \( h^t \). But, because Lemma 2 implies that both extremes in the inequality above converge to \( u(\theta) \), it thus follows that the term in the middle converges and has limit \( u(\theta) \). This result is formally stated in the following Proposition:

**Proposition 6** Let \( N \to \infty \). Then, \( \hat{u}_n \to u(\theta) \) for almost all histories.

Since the expected ability eventually converges to the truth, the effect of an additional signal converges to zero. Hence, the benefit of memory manipulation converges to zero and the agent does not engage in self deception after a sufficiently large amount of observations.

**Proposition 7** Let \( N \to \infty \). Then, \( r_H \to 0 \) and \( f_L \to 0 \) for almost all histories. Furthermore, under assumption 2b, there exists a (finite) \( \bar{N} \in \mathbb{N} \) such that \( n \geq \bar{N} \) implies \( r_H(h^n) = f_L(h^n) = 0 \) for almost all \( h^n \).

Denote by \( w(q; h^{n-1}) \) the weighting function associated with the utility of observing a signal in period \( n \) given history \( h^{n-1} \) and denote the coefficient of ambiguity aversion by \( c(q; u, h^n) \).

**Corollary 1** Let \( N \to \infty \). \( w(q, \cdot) \to q \) and \( c(q; u, \cdot) \to 0 \) for almost all histories. Furthermore, under assumption 2b, there exists a (finite) \( \bar{N} \in \mathbb{N} \) such that \( n \geq \bar{N} \) implies \( w(q; h^n) = q \) and \( c(q; u, h^n) = 0 \) for almost all \( h^n \).

Therefore, when signals are observed frequently enough, agents will not engage in self-deception and will not exhibit ambiguity aversion. 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).

### 4 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. In either case, the credit and the blame associated with an outcome depend, we suggest, on the attributions for success and failure. In the domain of chance, both success and failure are attributed primarily to luck. The situation is different when a person bets on his or her judgement. If the decision maker has limited understanding of the problem at hand, failure will be attributed to ignorance, whereas success is likely to be attributed to chance. In contrast, if the decision maker is an "expert," success is attributable to knowledge, whereas failure can sometimes be attributed to chance."

Heath and Tversky (1991, pp.7-8)

---

28Note that the probability of occurrence of a history \( \Pr(h^t) \) is a function of the sequence of memory manipulations \( r \) and \( f \). The previous assumption implies that the sets of histories with zero measure is the same for all relevant manipulation efforts: \( r_H(h^n) \in [0, 1 - \eta_H] \) and \( f_L(h^n) \in [0, f] \). Therefore, we omit any explicit reference to \( r \) and \( f \) when considering almost sure convergence of \( \hat{u}_n(h^n) \).
In this Section, the basic framework is generalized to allow for monetary lotteries. Consider a decision-maker (DM) with preferences over ability $\theta \in \Theta$ and money $t \in \mathbb{R}$ where, as before, the agent’s prior distribution of $\theta$ over the set $\Theta \subseteq \mathbb{R}$ has density $F(\cdot)$. Preferences are represented by a von Neumann-Morgenstern utility function $u : \Theta \times \mathbb{R} \to \mathbb{R}$ which is strictly increasing in both arguments.

The informational structure of the model is the following. Upon observing a signal $\sigma = s$, the DM receives a monetary payment $s \in \{L, H\}, L < H$. A high signal is assumed to be more favorable than a low signal in the sense of first-order stochastic dominance. The individual can manipulate her recollections $\hat{\sigma}$ by engaging in efforts $r_s$ and $f_s$, where $r_s \in [0, 1 - \eta_s]$ denotes the effort to remember and $f_s \in [0, \eta_s + r_s]$ denotes the effort to forget signal $s$. Memory manipulation $r_s$ and $f_s$ involves a cost of $\psi_r(r_s) \geq 0$ and $\psi_f(f_s) \geq 0$. Figure 7 depicts the extensive form representation of the game.

An important assumption is that the agent does not remember exactly how much money $t_0 \in \mathbb{R}$ she had before observing the signal. Therefore, she cannot fully infer which signal was observed from the amount of money she has in period 2.\footnote{Alternatively, we could assume that, by forgetting the outcome of the lottery, the agent also forgets the money she had before observing the signal.} Her beliefs about $t_0$ are given by the
absolutely continuous c.d.f. $G$ with strictly positive pdf $g$. Thus, the final monetary outcome is

$$t = t_0 + s,$$

where $s \in \{L, H\}$ the outcome of the lottery in period 1.

**Definition 3** A perfect Bayesian equilibrium (PBE) of the game is a strategy profile \( \{r_H^*(t), f_L^*(t)\}_{t \in \mathbb{R}} \) and posterior beliefs $\mu(\cdot | \hat{\sigma}, t)$ such that:

1. \( f_L^*(t) \in \arg \max_{f_L} (\eta_L - f_L) \int u(\theta, t) \, d\mu(\theta | \hat{\sigma} = L, t) \) + \( (1 - \eta_L + f_L) \int u(\theta, t) \, d\mu(\theta | \hat{\sigma} = \emptyset, t) \) - \( \psi_f(f_L) \)

2. \( r_H^*(t) \in \arg \max_{r_H} (\eta_H + r_H) \int u(\theta, t) \, d\mu(\theta | \hat{\sigma} = H, t) \) + \( (1 - \eta_H - r_H) \int u(\theta, t) \, d\mu(\theta | \hat{\sigma} = \emptyset, t) \) - \( \psi_r(r_H) \)

3. $\mu(\theta | \hat{\sigma} = x, t)$ is obtained by Bayes’ rule if $\Pr(\hat{\sigma} = x | f_L^*, r_H^*, t) > 0$.

A recollection $\hat{\sigma} \in \{L, H\}$ can only occur if the agent observed a signal $\sigma = \hat{\sigma}$ in date 1. Therefore, Bayesian updating implies that $\mu(\theta | \hat{\sigma}, t) = F(\theta | \sigma = \hat{\sigma})$ for $\hat{\sigma} \in \{L, H\}$ and payoffs are given by

$$u_L(t) = \int u(\theta, t) \, dF(\theta | \sigma = L), \quad \text{and} \quad u_H(t) = \int u(\theta, t) \, dF(\theta | \sigma = H).$$

When the DM forgets which signal was observed, there are 2 possible nodes: (i) $s = H$ and $t_0 = t - H$, and (ii) $s = L$ and $t_0 = t - L$. Bayesian updating gives

$$\Pr(\sigma = H | \hat{\sigma} = \emptyset, t) = \alpha(r_H, f_L, t),$$

where $\alpha(r_H, f_L, t) = \frac{q(1-r_H-\eta_H)q(t-H)}{q(1-r_H-\eta_H)q(t-H)+(1-q)(1-r_L-\eta_L)q(t-L)}$. Hence, the expected utility after forgetting a signal and having a monetary outcome $t$ is:

$$u_{\emptyset}(t) = \alpha(r_H^*(t), f_L^*(t), t) u_H(t) + \left[ 1 - \alpha(r_H^*(t), f_L^*(t), t) \right] u_L(t),$$

where $r_H^{*}(t)$ and $f_L^{*}(t)$ are the amount of memory manipulation that the date-2 self believes that was employed in period 1.

Proceeding as in Section 2, we obtain the following:

**Proposition 8** There exists a PBE. Suppose that $\psi_i$ is strictly convex, $i \in \{L, H\}$. Then, in any PBE we have:

1. If Assumption 2a holds, then $f_L^*(t) > 0, r_H^*(t) > 0 \forall t$.

2. If Assumption 2b holds, then

$$\psi'_f(0) \geq \Delta u(t) \implies f_L^*(t) = 0, \quad \text{and} \quad \psi'_r(0) \geq \Delta u(t) \implies r_H^*(t) = 0.$$
Proof. Analogous to Proposition 1. ■

Denote the distribution of \( t \) by \( h(t) \). Then, we have the following representation result:

**Proposition 9** The agent’s (ex-ante) expected utility from the lottery is

\[
U^\delta = MC + \Lambda,
\]

where:

- \( U^\delta \equiv \int \int u(\theta, t) f(\theta) h(t) dt d\theta \) (\( \theta \) and \( t \) independent),
- \( MC \equiv \int [q \psi_r(r_H^*(t)) + (1-q) \psi_f(f_L^*(t))] h(t) dt \), and
- \( \Lambda \equiv \int \xi(t) \Delta u(t) h(t) dt \),

\[
- \xi(t) \equiv [g(t - H) - g(t - L)] z(t), \text{ for some } z(t) > 0.
\]

The utility of a skill-dependent lottery can be decomposed into three terms. First, the utility \( U^\delta \) obtained from a lottery with the same distribution over monetary outcomes but whose outcomes are independent of \( \theta \). Second, the expected manipulation costs \( MC \). And, third, the degree of complementarity \( \Lambda \) between characteristics and the monetary outcomes. If the lottery is more likely to pay in states where the payoff gain from characteristics \( \Delta u(t) \) is high, then \( \Lambda \) is positive so that the lottery is more valuable. Otherwise, \( \Lambda \) is negative and the lottery is less valuable.

Denote the expected utility of receiving a signal \( s \in \{L, H\} \) by \( v_s \equiv \int u_i(t) h(t) dt \). The following theorem presents the nonexpected utility representation of the model.

**Theorem 2** The agent’s expected utility from the lottery can be represented by

\[
U(L) = w(q) v_H + [1 - w(q)] v_L,
\]

where

\[
w(q) = q + \frac{\Lambda - MC}{\int \Delta u(t) h(t) dt}.
\]

Furthermore,

i. \( w(0) = 0 \), and
ii. \( w(1) = 1 \).

**Example 4 (Exogenous Memory)** In the exogenous model of Example 3, we have \( w(q) = q + \frac{\Lambda}{E[\Delta u]} \). Therefore, the DM prefers a skill-dependent lottery to a lottery with the same distribution over monetary outcomes but whose monetary outcomes are uncorrelated with characteristics if and only if \( \Lambda \equiv E_t [\xi \times \Delta u] > 0 \).

\[30\]It can be shown that \( h(t) = q g(t - H) + (1-q) g(t - L), t \in \mathbb{R} \).
4.1 Discussion

The model presented here suggests that ambiguity aversion is a consequence of the lottery outcomes being informative about the decision-maker’s characteristics. When the decision-maker has an imperfect memory, she may exhibit ambiguity aversion/lovingness. Several experimental papers have argued that ambiguity aversion may be related to an agent’s skill or knowledge.31

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, those with graphic descriptions second, and those with 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.32

The idea that deviations from expected utility occur when lotteries are affected by the decision-maker’s skills dates back to Cohen and Hansel (1959) and Howell (1971). Howell considered composite gambles involving throwing a dart and spinning a roulette wheel. They varied the skill (dart) and chance (wheel) components of the lotteries while keeping the probability of winning constant. In general, people preferred lotteries with higher component of skill over chance. Cohen and Hansel presented a similar experiment where agents faced composite lotteries involving a mix of skill and chance components. The agents also tended to prefer lotteries with a higher component of skill over chance.

Heath and Tversky (1991) proposed the "competence hypothesis", according to which people’s preferences over ambiguous events arise from the anticipation of feeling knowledgeable or competent.33 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."

Fox and Tversky (1995) 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" argues that ambiguity aversion is driven by the feeling of incompetence. Fox and Weber (2002) replicated their results and extended them to the cases of familiar versus unfamiliar situations. Chow and Sarin (2001) obtained similar results.34

Similarly, Goodie (2003) proposed the perceived control hypothesis, according to which ambiguity aversion is generated by the agent’s belief that the distribution of outcomes is influenced

31 See Goodie and Young (2007) for a more detailed discussion of this literature.
32 See also Budescu et al. (2002).
34 In Chow and Sarin (2001), however, preference for risk decreased but did not completely disappear in non-comparative situations.
by characteristics such as knowledge or skill.\textsuperscript{35}

4.2 First Order Risk Aversion

In this subsection, it is shown that memory imperfection may lead the DM to exhibit first-order risk aversion [Segal and Spivak (1990)]. First-order risk aversion has important economic implications. An individual with second-order risk aversion always accepts small gambles with positive expected value.\textsuperscript{36} On the other hand, an individual with first-order risk aversion always rejects small gambles as long as the positive expected value is sufficiently small. As a consequence, someone with first-order risk aversion may choose to be fully insured even if prices are not actuarially fair.

Consider a lottery over money whose outcomes are informative about the decision maker’s characteristics as described previously. The certainty equivalent of the lottery is defined by the monetary amount $CE$ that makes the agent indifferent between participating in the lottery or receiving $CE$ for sure. In this case, receiving $CE$ for sure implies consuming $t_0 + CE$, which gives utility $\int_0 u(\theta, t_0 + CE) dF(\theta)$ for each realization $t_0$. Since $t_0$ is a random variable distributed according to a pdf $g$, the certainty equivalent is defined by the amount $CE$ such that

$$U(L) = \int_{t_0} \int_0 u(\theta, t_0 + CE) dF(\theta) g(t_0) dt_0,$$

where $U(L)$ is the expected utility from the lottery as described in Theorem 2. The risk premium associated to a lottery $L$ is defined as the difference between the expected payment and the certainty equivalent: $\pi \equiv E |t| - CE$.

Let $\tilde{t}$ be a binary random variable such that $E[\tilde{t}] = 0$. Consider a lottery that pays $t = \delta \tilde{t}$, where $\delta > 0$. We say that a decision maker has second-order risk aversion if $\lim_{\delta \to 0} \pi(\delta)/\delta = 0$, and we say that she has first-order risk aversion if $\lim_{\delta \to 0} \pi(\delta)/\delta > 0$.

First, let us consider the Exogenous Memory model of Example 3. In this case, there are no manipulation costs ($f^*_L = r^*_H = 0$) and the first-order risk attitude is determined solely by whether characteristics $\theta$ and money are complements or substitutes:

**Lemma 4** In the Exogenous Memory model:

- the DM is first-order risk-averse if $\int_t \Delta u(t) g'(t) dt > 0$,
- the DM is first-order risk-loving if $\int_t \Delta u(t) g'(t) dt < 0$,
- the DM has risk preference of second-order if $\int_t \Delta u(t) g'(t) dt = 0$.

The intuition follows from the representation of Proposition 9. In that proposition, we have decomposed the utility of a skill-dependent lottery in three terms: (i) the utility $U^I$ from a lottery with the same distribution over monetary outcomes but whose outcomes are independent of $\theta$, (ii) the expected manipulation costs $MC$, and (iii) the degree of complementarity $\Lambda$ between the characteristics $\theta$ and the monetary outcomes. Because an expected utility maximizer has risk preferences of second-order, the first term ($U^I$) does not influence her first-order risk preferences. Moreover, in the case of exogenous memory, the second term ($MC$) vanishes since there are no manipulation costs.

\textsuperscript{35}There is a large experimental litterature on the effect of perceived control on risk-taking [e.g. Chau and Phillips (1995) and Horswill and McKenna (1999)].

\textsuperscript{36}See Samuelson (1963), Segal and Spivak (1990), or Rabin (2000).
Therefore, in the exogenous memory model, first-order risk preferences are driven solely by the third term: the degree of complementarity between the agent’s characteristics \( \theta \) and the monetary outcomes. Taking the limit when \( H - L \to 0 \) implies that \( \Lambda > 0 \iff \int_t \Delta u(t) g'(t) dt < 0 \). Hence, a DM with exogenous memory is first-order risk-averse if characteristics and the monetary outcomes are substitutes \((g' < 0)\) and is first-order risk-loving if they are complements \((g' > 0)\).

When memory is endogenous, there is an additional term due to the costs of memory manipulation which reduces the utility of the lottery. Denote the equilibrium amounts of memory manipulation as a function of \( \delta \) by \( r^*_H(t; \delta) \) and \( f^*_L(t; \delta) \). It is useful to distinguish between two cases:

i. \( \Pr(\lim_{\delta \to 0} r^*_H(t; \delta) > 0) > 0 \) or \( \Pr(\lim_{\delta \to 0} f^*_L(t; \delta)) > 0 \), and

ii. \( \Pr(\lim_{\delta \to 0} r^*_H(t; \delta) > 0) = \Pr(\lim_{\delta \to 0} f^*_L(t; \delta) > 0) = 0 \),

where the probability is with respect to the distribution of \( t \). In case (i), the DM engages in memory manipulation in some set with positive measure even when the monetary outcomes from the lottery are arbitrarily small. This would always be the case if the informational content of the lottery \( \Delta u(t) \) is large enough (in some set with positive measure) or if the marginal cost of manipulation \( \psi'_f(0) \) is small enough. In case (ii), the amount of memory manipulation converges to zero almost surely.

In case (i), the DM would demand a participation premium in order to observe the signal when \( \delta = 0 \). Therefore, the certainty equivalent of the lottery converges to \( CE(0) < 0 \) and

\[
\lim_{\delta \to 0} \frac{\pi(\delta)}{\delta} = - \lim_{\delta \to 0} \frac{CE(\delta)}{\delta} = +\infty.
\]

Thus, when manipulation costs do not converge to zero, they dominate the complementarity effect and the DM always exhibits first-order risk aversion.

In case (ii), the answer depends on which effect dominates. If characteristics \( \theta \) and the monetary outcomes are complements \((g' > 0)\), then both push towards risk-aversion and the DM exhibits first-order risk aversion. If \( \theta \) and the monetary outcomes are substitutes \((g' < 0)\), then the behavior of DM depends on the rate of convergence of manipulation costs compared to the complementarity between \( \theta \) and the monetary outcomes.

These results are formally stated in the lemma below:

**Lemma 5** Suppose \( \Pr(\lim_{\delta \to 0} r^*_H(t; \delta) > 0) > 0 \) or \( \Pr(\lim_{\delta \to 0} f^*_L(t; \delta)) > 0 \). Then, \( CE(0) < 0 \) and DM displays first-order risk aversion. Suppose \( \Pr(\lim_{\delta \to 0} r^*_H(t; \delta) > 0) = \Pr(\lim_{\delta \to 0} f^*_L(t; \delta)) = 0 \). Then, there exists some strictly increasing \( \varpi(q) > 0 \) such that

1. DM is first-order risk averse if

\[
\varpi(q) \int_t \Delta u(t) g'(t) dt + \int_t [q\psi'_r(r_H(t; 0)) r'_H(0) + (1 - q) \psi'_f(f_L(t; \delta)) f'_L(0)] g(t) dt > 0,
\]

and

2. DM is first-order risk loving if

\[
\varpi(q) \int_t \Delta u(t) g'(t) dt + \int_t [q\psi'_r(r_H(t; 0)) r'_H(0) + (1 - q) \psi'_f(f_L(t; \delta)) f'_L(0)] g(t) dt < 0.
\]
Of course, when there are no complementarities between skill and money (e.g., when \( g \) is uniform) and when \( \Pr (\lim_{\delta \to 0} r_H^* (t; \delta) > 0) = \Pr (\lim_{\delta \to 0} f_L^* (t; \delta)) = 0 \), then DM exhibits second-order risk preferences.

Lemma 5 relies on assumptions made on the endogenous variables \( r_H \) and \( f_L \). Next, we present some results based on properties of the fundamentals \( \psi \) and \( \Delta u (t) \):

**Proposition 10** Under Assumption 2a, DM exhibits first-order risk aversion. Under Assumption 2b, there exist \( \lambda \) and \( \hat{\lambda} \) with \( \lambda \geq \hat{\lambda} > 0 \) such that, in any PBE,

- DM is first-order risk averse if \( \Pr (\Delta u (t) > \hat{\lambda}) > 0 \),
- DM is first-order risk loving if \( \Pr (\Delta u (t) > \lambda) = 0 \) and \( \int_t \Delta u (t) g' (t) dt < 0 \).

**Proof.** Follows from Lemmas 1 and 5. □

**Remark 2** Consider a repeated version of the game above. Proceeding as in Section 3, it follows that \( \Delta u_n (t) \to 0 \) for almost all histories. Therefore, the DM’s attitude towards risk converges to second-order risk aversion as \( N \to \infty \) (for almost all histories).

### 4.3 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.\(^{37}\) In this subsection, I provide 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, the outcome of trading 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 requires a participation premium in order to engage in an activity that reveals information about her skills. Therefore, she may prefer not to trade if the price is only slightly above the expected value of the good.

The model presented below establishes this result formally. It consists of a special case of the model from Section 4. The main difference from the original model is in the particular specification of the signal \( \sigma \) observed in period 1.

A decision-maker owns good \( A \) and must choose whether or not to trade it for good \( B \). Good \( A \) has a known monetary value of \( v_A \in \mathbb{R} \). The value of good \( B \) is unknown. It may take values \( v_B \in \{ v_L, v_H \} \), where \( v_L < v_A < v_H \).

The DM must evaluate whether trade is profitable. This is formalized by a signal \( s \in \{ T, NT \} \) denoting whether the agent should or should not trade. This signal is informative about the

\(^{37}\) According to loss aversion, losses are weighed substantially more than gains. Then, the cost of losing a good that is much higher than the benefit of winning a good.
probability of trade being profitable. Predicting the realization of $B$ correctly requires skills. Therefore, a correct prediction is good news about DM’s skills. However, $s$ is only informative about the DM’s skills through the realization of each state ($v^L$ and $v^H$). More specifically, $s$ is uninformative about $\theta$ if $v_B$ is not observed. This information structure is formally stated below:

**Assumptions** Let $(\theta, v_B, s) \in \Theta \times \{v^L, v^H\} \times \{T, NT\}$ be distributed according to the c.d.f. $F(\theta, v_B, s)$.

2.1 $\theta$ and $s$ are independent:

$$F(\theta, s = \hat{s}) = F(\theta) \times \Pr(s = \hat{s}), \quad \text{for} \; \hat{s} \in \{T, NT\}.$$ 

2.2 The signal $s \in \{T, NT\}$ is informative about the value of good $B$, $v^B$:

$$q^H \equiv \Pr(v_B = v^H | s = T) > \Pr(v_B = v^H) \equiv \mu,$$

$$q^L \equiv \Pr(v_B = v^L | s = NT) > 1 - \mu.$$ 

2.3 The signal $s \in \{T, NT\}$ is informative about $\theta$ when the realization of $v_B$ is observed:

$$F(\theta | v_B = v^H, s = T) \leq F(\theta | v_B = v^H, s = NT),$$

$$F(\theta | v_B = v^L, s = NT) \leq F(\theta | v_B = v^H, s = T),$$

which strict inequality over a set with non-zero measure.

2.4 It is profitable to trade if and only if $s = T$.

$$q^H v^H + (1 - q^H) v^L > v_A > \mu v^H + (1 - \mu) v^L.$$ 

Note that Assumption 2.1 implies that the signal $s$ is only informative about $\theta$ if the value of good $B$ is also observed:

$$F(\theta | s = NT) = F(\theta | s = T) = F(\theta). \quad (14)$$

Assumption 2.4 states that it is profitable to trade conditional on an $s = T$ signal and that it is not profitable to trade if no signal is observed. It implies that it is also not profitable to trade conditional on an $s = NT$ signal (since $s = NT$ is bad news about the profitability of trade).

The timing of the model is as follows:

1. Nature determines $\theta$, the trading skills (or knowledge) of the DM. However, this is not observed by the agent.

2. The DM observes a signal $s$, which is correlated with the benefits from trade but, by itself, uninformative about her skills. Then, she chooses whether to trade good $A$ for good $B$.

3. If the DM chose to trade, she observes $v_B$, the value of good $B$. The value of good $B$ along with signal $s$ is informative about her skills $\theta$.

4. The DM exerts efforts $r_H$ and $f_L$ in order to forget unsuccessful and remember successful outcomes. The effort costs are given by $\psi_L(f_L)$ and $\psi_H(r_H)$, respectively.
For simplicity, I assume that \( t_0 \) is uniformly distributed over a large interval so that there are no complementarities between ability and money (\( \Lambda = 0 \)).\(^{38}\) Therefore, Proposition 9 implies that the agent always prefers a lottery whose outcomes are uncorrelated with skills to one with the same distribution over monetary outcomes but where outcomes are correlated with her skills. Furthermore, I also assume that she is risk-neutral so that the von Neumann-Morgenstern utility function can be written as \( u (\theta, t) = \phi (\theta) \times t \) for some strictly increasing function \( \phi \) such that \( \int \phi (\theta) f (\theta) d\theta = 1.\(^{39}\)

Upon observing a no-trade signal, trading leads to a lottery whose monetary outcomes are correlated with skills. Since \( \Lambda = 0 \), the expected utility of the lottery is equal to the expected monetary payoffs \( q^L v^L + (1 - q^L) v^H - v_A \) minus the expected manipulation costs. Manipulation costs are always weakly negative and, when \( s = NT \), the expected monetary payoffs are strictly negative. Therefore, the agent prefers not to trade when she observes \( s = NT \).

Consider the case where the DM observes a trade signal and suppose she chooses to trade the good. If good \( B \) turns out to have a high value \( v_B = v^H \) so that the prediction was correct, then her posterior over \( \theta \) becomes \( F (\theta | v_B = v^H, s = T) \). Letting \( t = t_0 + v^H \) where \( t_0 \) is the (unknown) initial amount of money as in Section 4, the expected payoff conditional on \( v_B = v^H \) becomes

\[
 u_H (t) = t \times \int \phi (\theta) dF (\theta | v_B = v^H, s = T). 
\]

If good \( B \) turns out to have a low value, her expected payoff becomes

\[
 u_L (t) = t \times \int \phi (\theta) dF (\theta | v_B = v^H, s = T). 
\]

Let \( \Delta \phi \equiv \int \phi (\theta) dF (\theta | v_B = v^H, s = T) - \int \phi (\theta) dF (\theta | v_B = v^H, s = T) > 0 \). Then, \( \Delta u (t) = t \times \Delta \phi \).

If the DM forgets the signal, there are 2 possibilities. Trade may have been successful but she had a low initial endowment \( v_B = v^H, t_0 = t - H \), or trade may have been unsuccessful but she had a high initial endowment \( v_B = v^L, t_0 = t - L \). Bayesian updating yields

\[
 u_{\phi} (t) = \alpha (r_H^* (t), f_L^* (t)) u_H (t) + [1 - \alpha (r_H^* (t), f_L^* (t))] u_L (t),
\]

where \( \alpha (r_H, f_L) = \frac{q^H v^H + (1 - q^H) v^L}{q^H + (1 - q^H) v^L} \) is the Bayesian weight.\(^{40}\)

From Proposition 9, the utility from trading conditional on observing a signal \( s = T \) is

\[
 u_T (t) = \mathbb{E} \left[ \begin{array}{c}
 q^H v^H + (1 - q^H) v^L - MC
 \end{array} \right],
\]

whereas the utility from not-trading is \( u^A \). Since \( q^H v^H + (1 - q^H) v^L > v^A \), DM prefers to trade if and only if manipulation costs are "not too high". If DM could commit not to engage in self-deception, she would always trade. Hence, self-deception implies that some trades with positive expected monetary gains do not occur. The agent balances a positive expected gain from trade with the costs of self-deception.

\(^{38}\) Let \( t_0 \) be uniformly distributed over \([0, K]\) for \( K \) large. When \( t \in [L, H) \), DM infers that \( s = L \) regardless of her recollections \( \hat{\sigma} \) and, therefore, does not engage in memory manipulation. Similarly, when \( t \in (K + L, K + H] \), she infers that \( s = H \) and does not exert memory manipulation. When \( K \) is large, however, these intervals happen with small probability. When \( t \in [H, K + L) \), the solution is obtained as in Section 4.

\(^{39}\) The normalization \( \int \phi (\theta) f (\theta) d\theta = 1 \) implies that the expected utility of a lottery with expected monetary payoff equal to \( t \) and whose monetary outcomes are independent of skills is equal to \( t \).

\(^{40}\) Note that \( \hat{\alpha} \) is not a function of \( t \) because we have assumed that \( t_0 \) is uniformly distributed.
We say that the agent displays an endowment effect if she requires an expected value of trade strictly higher than her valuation of the good in order to trade. In this model, the agent displays an endowment effect whenever \( MC > 0 \) since she is only willing to trade if the expected value of \( B \) is strictly greater than the value of \( A \). The result is summarized in the following proposition:

**Proposition 11** The agent will refuse trades with positive expected gains if

- self-deception is sufficiently costly (i.e. \( MC \) is high enough), or
- the expected value of trading is sufficiently small (i.e. \( q^Hv^H + (1 - q^H)v^L - v^A \) is small enough).

Furthermore, under Assumption 2a, there is an endowment effect.

**Proof.** Analogous to Propositions 8 and 10. Note that \( \lambda / \Delta \varphi \) and \( \hat{\lambda} / \Delta \varphi \) are such that, for any finite \( \lambda \), \( \Pr \left( t > \frac{\lambda}{\Delta \varphi} \right) > 0 \). ■

**Remark 3** Consider a repeated setting as in Section 3. Because memory manipulation converges to zero after the individual has played the lottery a sufficiently large number of times, the model implies that people who trade often enough should not exhibit an endowment effect. This result is consistent with evidence from List (2003).

### 4.4 The Uncertainty Effect

In a set of experiments with over 1000 participants, Gneezy, List, and Wu (2006) found that a significant proportion of participants seemed to value participating in a lottery less than obtaining the lowest possible outcome for sure. For example, they were willing to pay $38 for a $50 gift certificate but only $28 for a lottery ticket that provided an equal chance of winning the same $50 certificate and a $100 certificate with the same conditions. As Gneezy, List, and Wu argue, “[t]his behavioral result, which we term the uncertainty effect, not only contradicts expected utility and prospect theory but is inconsistent with virtually all models of risky choice.”

The present model provides a memory manipulation explanation for the uncertainty effect. If the lottery outcomes lead the decision-maker to exert memory manipulation, she may prefer to obtain the worst monetary outcome for sure and avoid the additional memory costs.

If DM receives the lowest possible monetary outcome \( (L) \) for sure, the lottery is not informative about her characteristics. Thus, she obtains an expected payoff equal to

\[
\int_{t_0} \int_{\theta} u(\theta, t_0 + L) g(t_0) d\mathbf{F}(\theta) dt_0.
\]

However, by participating in the lottery, the individual may infer something about her characteristics from the outcome. Therefore, from Proposition 9, her expected utility from the lottery is

\[
\int_{t_0} \int_{\theta} u(\theta, t) h(t) d\mathbf{F}(\theta) dt - MC + \Lambda.
\]

Whether obtaining \( t_L \) for sure is preferred to participating in the lottery depends on the expected monetary gain from the possibility of winning a higher payment under the lottery, the expected self-deception costs, and the complementarity between characteristics and money. Whenever the expected monetary gains and the complementarity \( \Lambda \) between \( \theta \) and money are lower than the expected memory costs \( MC \), she will prefer to obtain \( L \) for sure:
Lemma 6 A decision-maker prefers to obtain the lowest monetary outcome for sure instead of participating in the lottery if and only if
\[ \int_{t_0}^{t} u(\theta, t_0 + L) g(t_0) dF(\theta) dt_0 \leq \int_{t}^{t} u(\theta, t) h(t) dF(\theta) dt - MC + \Lambda. \] (16)

Suppose the DM is risk neutral so that \( u(\theta, t) = \phi(\theta) \times t \) for some function \( \phi \) with \( \int \phi(\theta) f(\theta) d\theta = 1 \). The, the expected monetary gain of participating in the lottery is \( q\Delta t \), where \( \Delta t = H - L \).

Corollary 2 Suppose the DM is risk neutral. Then, she prefers to obtain the lowest monetary outcome for sure instead of participating in a lottery whose outcomes are informative about \( \theta \) if and only if the expected monetary gains plus the gains from complementarity are greater than the expected memory manipulation costs:
\[ MC \leq q\Delta t + \Lambda. \] (17)

We say that DM displays an uncertainty effect if she prefers receiving \( L \) for sure instead of participating in a lottery that pays \( L + \Delta t \) with probability \( q \) and \( L \) with probability \( 1 - q \) for some \( \Delta t > 0 \). As in Subsection 4.3, let \( t_0 \) be uniformly distributed over a large interval so that \( \Lambda = 0 \). Condition (17) implies that, whenever DM engages in memory manipulation, there will be an interval \((0, \Delta t)\) such that receiving \( L \) for sure is preferred to participating in the lottery:

Proposition 12 Suppose DM is risk neutral and let \( t_0 \) be uniformly distributed over a large interval. Under Assumption 2a, DM displays an uncertainty effect. Furthermore, there exists a \( \Delta \bar{t} > 0 \) such that the lowest monetary outcome is preferred to the lottery if \( \Delta t < \Delta \bar{t} \).

Because the DM requires a participation premium in order to accept to participate in a lottery that is informative about \( \theta \), she will prefer to receive the lowest possible monetary payoff for sure if the expected payments from the lottery are low enough.

4.5 The Sunk Cost Fallacy

“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.\(^{41}\)

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.

\(^{41}\)Sunk costs effects are also called “irrational escalation of commitment”, the “entrapment effect”, or “too much invested to quit”.

31
Camerer and Weber (1999) analyzed the market for professional basketball players. Players who are drafted earlier represent larger sunk costs. They show that, conditional on performance, players who are drafted earlier get more playing time, which suggests the presence of sunk cost effects.

In this subsection, I present a self-deception model of 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”.

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 to sunk monetary costs but real psychological costs. Abandoning a project usually involves admitting that a wrong decision was made. Therefore, revising one’s position in the project reveals information about her skills or knowledge. As shown in Subsection 2.5, the DM requires a participation premium in order to engage in an activity that reveals information about her characteristics. Hence, some projects with negative expected value will not be terminated. This subsection incorporates the basic model of Section 4 into a reversible investment model.

Consider a decision-maker that has an opportunity invest in a project, which involves a sunk cost equal to $K > 0$. The project will give a monetary payoff equal to $t \in \{t^1, t^2, \ldots, t^N\}$, $t^1 < t^2 < \ldots < t^N$, $N \geq 3$, which is not known ex-ante. The monetary outcome of the project is informative about the DM’s skills. Higher monetary payoffs are better news about the DM’s skills:

$$F(\theta | t = t^{n+1}) \leq F(\theta | t = t^n) \text{ for all } \theta \in \Theta,$$

with strict inequality in some set with positive measure for all $n \in \{1, \ldots, N-1\}$.

After the sunk investment was made, the decision-maker can reevaluate the value of the project. More precisely, she can choose whether to observe the value of the project $t$ (at zero cost). She also decides whether to terminate it. Subsequently, she can manipulate her recollections of the profitability of the project. If the project was not terminated, the DM receives a monetary payment of $t$.

As in Subsection 4.3, I will assume that $t_0$ is uniformly distributed over a large interval ($A = 0$). Furthermore, the DM is assumed to be risk-neutral so that $u(\theta, t) = \phi(\theta) \times t$ for some function $\phi$ with $\int \phi(\theta) f(\theta) d\theta = 1$.

The project is ex-ante efficient but fails in some cases

$$E[t] \geq K, \text{ and}$$

$$t^2 < 0 < t^3.$$

Therefore, the project fails if $t \in \{t^1, t^2\}$. Denote the expected payoff from skills conditional on $t^i$ by $\phi_i \equiv \int \phi(\theta) dF(\theta | t = t^i) dt$. Then, the agent’s utility given $t = t^i$ is $\phi_i \times [t^i + E(t_0)]$.

The timing of the model is as follows:

1. The DM chooses whether or not to invest in the project. Investing costs $K > 0$.

2. If the investment was made, she decides whether or not to reassess the profitability of the project, which takes value $t^i$ with probability $q_i \in (0, 1), i = 1, \ldots, N$.

See Brockner (1992) for a review of the literature.
3. The DM chooses whether or not to abort the project.

4. If the project is aborted (A), the DM remembers the signal with probability \( \eta_i + r_i - f_i \), where \( \eta_i \) is the natural rate of remembering signal \( i \), \( r_i \) is the amount of effort to remember and \( f_i \) is the effort to forget signal \( i \). If it is not aborted (NA), the agent observes \( t_i \) and obtains utility \( \phi_i \times \left[ t_i + E(t_0) \right] \).

The main idea is that abandoning the project is informative about the agent’s skills, which leads to self-deception. In this model, memory manipulation is not effective when the project is not aborted since the agent will observe of the project profitability in the next period anyway. However, the DM may want to manipulate her memory when the project is aborted in order to convince herself that her skills or knowledge are not extremely low even though the project was abandoned. By not reassessing the project after the sunk investment was made, the DM avoids the self-deception costs that follow termination. Therefore, she would prefer not to reassess the project if the value of information is lower than the costs of memory manipulation. In that case, however, inefficient projects \( t \in \{t_1, t_2\} \) are not aborted.

Suppose the DM chooses to observe the profitability of the project. Note that she will never terminate the project if she observes \( t^N \) since continuing gives a payoff of \( \phi_N \times \left[ t^N + E(t_0) \right] \), which is strictly greater than any payoff she could get under termination (both \( \phi_i \) and \( t^i \) are increasing in \( i \) and manipulation costs are nonnegative). Applying this reasoning inductively, it follows that the agent will never terminate a project if she observes \( t^i \), \( i \geq 3 \).

**Lemma 7** Suppose \( \text{Terminate} \neq \emptyset \). Then, in any PBE, the decision-maker never terminates a project after observing \( t^i \) such that \( i \geq 3 \).

Now consider a project \( i \in \{1, 2\} \). Terminating yields a payoff of

\[
[(\eta_i + r_i - f_i) \phi_i + (1 - \eta_i - r_i + f_i) E(\phi|\text{Terminate})] \times E(t_0) - MC_i,
\]

where \( \text{Terminate} \subseteq \{1, 2\} \) is the set of projects that the agent terminates and \( MC_i = \psi_L(f_i) + \psi_H(r_i) \) is the manipulation cost after observing \( t^i \). Continuing the project yields \( \phi_i \times \left[ t^i + E(t_0) \right] \).

The agent terminates a project after observing \( t^i \) if

\[
-\phi_i t^i \geq (1 - \eta - r_i + f_i) [\phi_i - E(\phi|\text{Terminate})] \times E(t_0) + MC_i.
\]

Hence, the DM continues project \( i \) if and only if the monetary loss from continuing the inefficient project \( (-\phi_i t^i > 0) \) is lower than the gain from remembering that the project was not as bad as the other projects which are interrupted \( (1 - \eta - r_i + f_i) [\phi_i - E(\phi|\text{Terminate})] \times E(t_0) \) plus the costs of memory manipulation \( MC_i \) that would be exerted if the project was interrupted.

**Proposition 13** Suppose Assumption 2a holds. There exist \( \hat{t}_1 < 0 \) and \( \hat{t}_2 < 0 \) such that if either \( t_1 > \hat{t}_1 \) or \( t_2 > \hat{t}_2 \), then the decision-maker does not terminate some inefficient projects in any PBE.

Thus, when the monetary losses \( (t_1 \text{ and } t_2) \) from continuing an inefficient project are "not too large", the DM will not terminate it. Although the investment involves sunk monetary costs, abandoning inefficient projects involve admitting one’s failure which leads to additional psychological costs.
5 Conclusion

This paper proposed a non-expected utility model based on self-deception. The model assumes that people have preferences over their perceived characteristics and that they can, to some extent, manipulate their memories. However, because self-deception is costly, an individual may prefer to avoid signals that are informative about her skills.

The predicted behavior is similar to that of a decision-maker with preferences over ambiguous events. However, since the non-linearity arises from the anticipation of self-deception costs, it predicts that individuals will avoid the events that lead to more self-deception. The model is consistent with recent experimental evidence that suggests that ambiguity aversion/lovingness is related to the decision-maker’s skills or knowledge. It also leads to a theory of ego utility and information acquisition.

It was shown that self-deception provides a unified explanation for several biases in decision-making. I have argued that the endowment effect can arise from an agent’s fear of learning that the trade turned out to be unprofitable. Similarly, sunk costs may influence behavior because a decision-maker may not want to admit that she made an unprofitable investment. In cases where self-deception is severe or the potential monetary gains are small, the decision-maker may even prefer to obtain the smallest monetary outcome for sure instead of participating in a lottery (i.e., she exhibits an uncertainty effect).

Because the decision-maker is assumed to be a Bayesian, she eventually learns the truth in an infinitely repeated environment. Hence, the model implies that her behavior converges to that implied by expected utility theory as the decision-maker gains experience.

Appendix

A Proofs

Proof of the claim in Remark 1:

Let \( \hat{\mu} \) and \( \mu \) denote the cumulative distribution functions of \( \hat{\theta}_\sigma \) and \( \theta_\sigma \), respectively. \( \hat{\theta}_\sigma \) second-order stochastically dominates \( \theta_\sigma \) if and only if, for any concave function \( g : \Theta \to \mathbb{R} \),

\[
\int g(\hat{\theta}_\sigma) \, d\mu(\hat{\theta}_\sigma) \geq \int g(\theta_\sigma) \, d\mu(\theta_\sigma).
\]

(20)

But

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

\[
\int g(\hat{\theta}_\sigma) \, d\mu(\hat{\theta}_\sigma) = q(r_H + \eta_H)g(\theta_H) + [q(1 - r_H - \eta_H) + (1 - q)(1 + f_L - \eta_L)]g(\hat{\theta}_\sigma) + (1 - q)(\eta_L - f_L)g(\theta_L).
\]

Substituting in inequality (20), yields

\[
[q(1 - r_H - \eta_H) + (1 - q)(1 + f_L - \eta_L)]g(\hat{\theta}_\sigma) > q(1 - r_H - \eta_H)g(\theta_H) + (1 - q)(1 - \eta_L + f_L)g(\theta_L).
\]
Because we have a contradiction. Therefore, suppose suppose

\[ f \]

which is true because \( g \) is concave.

Proof of Proposition 1:

Define \( \hat{f}_L (f_L, r_H) \) and \( \hat{r}_H (f_L, r_H) \) as the set of maxima of (2) and (3), respectively (i.e., these are the best-response correspondences of the date-1 selves). Since (2) and (3) are continuous and concave functions defined over a compact set, \( \hat{f}_L \) and \( \hat{r}_H \) are non-empty, convex, and compact sets. Define the transformation \( T : [0, \eta_L] \times [0, 1 - \eta_H] \to [0, \eta_L] \times [0, 1 - \eta_H] \) by \( T (f_L, r_H) = \left( \hat{f}_L (f_L, r_H), \hat{r}_H (f_L, r_H) \right) \). Then, Kakutani’s theorem establishes that there exists a fixed-point of \( T \), which is a PBE.

The other claims follow from the Kuhn-Tucker conditions.

1. For any \( f_L \in \hat{f}_L (f_L^*, r_H^*) \) and \( r_H \in \hat{r}_H (f_L^*, r_H^*) \),

\[
\alpha (f_L^*, r_H^*) \Delta u > 0 \implies f_L > 0, \quad \text{and} \quad [1 - \alpha (f_L^*, r_H^*)] \Delta u > 0 \implies r_H > 0.
\]

Suppose \( f_L^* = 0 \). Then, we must have \( \alpha (0, \eta_H^*) \Delta u \leq 0 = \psi_f (0) \). If \( \alpha (0, \eta_H^*) > 0 \), we already have a contradiction. Therefore, suppose \( \alpha (0, \eta_H^*) = 0 \). But \( \alpha (0, \eta_H^*) = 0 \iff 1 - \eta_H = r_H \), which is strictly positive since \( \eta_H < 1 \).

Now suppose \( r_H^* = 0 \). Then, we must have \( [1 - \alpha (f_L^*, \eta_H^*)] \Delta u \leq 0 \). As shown above \( f_L > 0 \). But, \( f_L > 0 \) implies \( 1 - \alpha (f_L, 0) > 0 \), which establishes that \( [1 - \alpha (f_L^*, \eta_H^*)] \Delta u > 0 \) so that \( r_H^* > 0 \).

2. For any \( f_L \in \hat{f}_L (f_L^*, r_H^*) \) and \( r_H \in \hat{r}_H (f_L^*, r_H^*) \),

\[
\alpha (f_L^*, r_H^*) \Delta u \leq \psi_f (0) \iff f_L = 0, \quad \text{and} \quad [1 - \alpha (f_L^*, r_H^*)] \Delta u \leq \psi_r (0) \iff r_H = 0.
\]

Since \( \alpha (f_L^*, r_H^*) \in [0, 1] \), it follows that \( \hat{f}_L (f_L^*, r_H^*) = \hat{r}_H (f_L^*, r_H^*) = \{0\} \).

Proof of Lemma 1:

There exists a PBE with \( f_L^* = r_H^* = 0 \) if

\[
[1 - \alpha (0, 0)] \Delta u \leq \psi_r (0), \quad \text{and} \quad \alpha (0, 0) \Delta u \leq \psi_f (0).
\]

Because \( \alpha (0, 0) = \frac{q (1 - \eta_L)}{q (1 - \eta_H) + (1 - q) (1 - \eta_L)} \), letting

\[
\hat{\lambda} = \min \left\{ \psi_r (0) \left[ 1 + \frac{q (1 - \eta_H)}{(1 - q) (1 - \eta_L)} \right]; \psi_f (0) \left[ 1 + \frac{(1 - q) (1 - \eta_L)}{q (1 - \eta_H)} \right] \right\}
\]

establishes that \( \Delta u \leq \hat{\lambda} \) implies that there exists a PBE with \( f_L^* = r_H^* = 0 \).

Next, we need to show that there is no PBE with \( \max \{f_L^*, r_H^*\} > 0 \) when \( \Delta u < \hat{\lambda} \). The proof is obtained by contradiction. Suppose that we have a PBE such that \( \max \{f_L^*, r_H^*\} > 0 \). There are 8 possible cases:
Case 1) $f^*_L = \eta_L, r^*_H = 1 - \eta_H$:
From the Kuhn-Tucker conditions, this is an equilibrium if

$$0 \geq \psi'_f(\eta_L), \text{ and } \Delta u \geq \psi'_r(1 - \eta_H).$$

Hence, $\psi'_f(\eta_L) > 0$ implies that we cannot have a PBE with $f^*_L = \eta_L, r^*_H = 1 - \eta_H$.

Case 2) $f^*_L = 0, r^*_H = 1 - \eta_H$:
This is a PBE if the following conditions hold:

$$\alpha (0, 1 - \eta_H) \Delta u \leq \psi'_f(0),$$

$$[1 - \alpha (0, 1 - \eta_H)] \Delta u \geq \psi'_r(1 - \eta_H).$$

Using the fact that $\alpha (0, 1 - \eta_H) = 0$, it follows that the first inequality always hold (from assumption 2b). Thus, we have a PBE with $f^*_L = 0, r^*_H = 1 - \eta_H$ if $\Delta u \geq \psi'_r(1 - \eta_H)$.

Case 3) $f^*_L \in (0, \eta_L), r^*_H = 1 - \eta_H$:
Using Kuhn-Tucker’s conditions and the fact that $\alpha (f^*_L, 1 - \eta_H) = 0$, it follows that $\psi'_f(f^*_L) = 0$, which is a contradiction. Hence, we cannot have a PBE with $f^*_L \in (0, \eta_L), r^*_H = 1 - \eta_H$.

Case 4) $f^*_L = \eta_L, r^*_H = 0$:
Kuhn-Tucker’s conditions give

$$\alpha (f^*_L, r^*_H) \Delta u \geq \psi'_f(\eta_L),$$

$$[1 - \alpha (f^*_L, r^*_H)] \Delta u \leq \psi'_r(0).$$

Because $\alpha (\eta_L, r_H) = \frac{q(1 - \eta_H - r_H)}{q(1 - \eta_H - r_H) + 1 - q}$, it follows that this is a PBE if:

$$\left[1 + \frac{q}{1 - q} (1 - \eta_H) \right] \psi'_r(0) \geq \Delta u \geq \left[1 + \frac{1 - q}{q(1 - \eta_H)} \right] \psi'_f(\eta_L).$$

Case 5) $f^*_L = \eta_L, r^*_H \in (0, 1 - \eta_H)$:
Kuhn-Tucker’s conditions imply that this is a PBE if

$$\left[1 + \frac{q(1 - \eta_H - r^*_H)}{1 - q} \right] \psi'_r(r^*_H) = \Delta u \geq \left[1 + \frac{1 - q}{q(1 - \eta_H - r_H)} \right] \psi'_f(\eta_L),$$

for some $r^*_H \in (0, 1 - \eta_H)$.

Case 6) $f^*_L \in (0, \eta_L), r^*_H \in (0, 1 - \eta_H)$:
This is a PBE if

$$\Delta u = \psi'_f(f^*_L) \left[1 + \frac{(1 - q)(1 - \eta_L + f^*_L)}{q(1 - \eta_H - r^*_H)} \right] = \psi'_r(r^*_H) \left[1 + \frac{q(1 - \eta_H - r_H)}{(1 - q)(1 - \eta_L + f^*_L)} \right],$$

for some $r^*_H \in (0, 1 - \eta_H)$, and some $f^*_L \in (0, \eta_L)$.

Case 7) $f^*_L = 0, r^*_H \in (0, 1 - \eta_H)$:
Kuhn-Tucker’s conditions yield

$$\psi'_r(r^*_H) \left[1 + \frac{q(1 - \eta_H - r_H)}{(1 - q)(1 - \eta_L)} \right] = \Delta u \leq \psi'_f(0) \left[1 + \frac{(1 - q)(1 - \eta_L)}{q(1 - \eta_H - r_H)} \right],$$

36
for some $r_H \in (0, 1 - \eta_H)$.

Case 8) $f_L^* \in (0, \eta_L)$, $r_L^* = 0$:
This is a PBE if

$$\psi_f'(f_L) \left[ 1 + \frac{(1 - q)(1 - \eta_L + f_L)}{q(1 - \eta_H)} \right] = \Delta u \leq \psi_f'(0) \left[ 1 + \frac{q(1 - \eta_H)}{(1 - q)(1 - \eta_L + f_L)} \right],$$

for some $f_L \in (0, \eta_L)$.

It is immediate that we cannot have a PBE in cases 1 and 3. Define $\lambda_1$ as

$$\min \left\{ \psi'_r(1 - \eta_H); \left[ 1 + \frac{1 - q}{q(1 - \eta_H)} \right] \psi'_f(\eta_L) \right\}.$$

Then, when $\Delta u < \lambda_1$, there are no PBE in cases 2 and 4.

Note that $\inf_{r_H} \left\{ \left[ 1 + \frac{1 - q}{q(1 - \eta_H - r_H)} \right] \psi'_f(\eta_L) \right\} > 0$, $\inf_{r_H, f_L} \left\{ \psi'_r(r_H) \left[ 1 + \frac{q(1 - \eta_H - r_H)}{(1 - q)(1 - \eta_L + f_L)} \right] \right\} > 0$, and $\inf_{r_H, f_L} \left\{ \psi'_f(f_L) \left[ 1 + \frac{(1 - q)(1 - \eta_L + f_L)}{q(1 - \eta_H - r_H)} \right] \right\} > 0$. Define $\lambda_2 > 0$ as the minimum of these three terms. Then, when $\Delta u < \lambda_2$, it follows that the conditions to Cases 5, 6, 7, and 8 cannot be satisfied. Hence, letting $\lambda = \min \{ \lambda_1, \lambda_2 \}$ establishes that we cannot have a PBE with $\max \{ f_L^*, r_L^* \} > 0$ when $\Delta u < \lambda$, which concludes the proof.  

Proof of Proposition 2:

First, notice that existence follows from Proposition 1. The characterization follows from Kuhn-Tucker’s conditions:

$$\Delta u \leq \psi_f'(0) \implies f_L = 0,$$

$$\Delta u \geq \psi_f'(1) \implies f_L = 1,$$

$$\left( 1 + \frac{1 - q}{q} f_L \right) \psi_f'(0) \leq \Delta u \leq \left( 1 + \frac{1 - q}{q} f_L \right) \psi_f'(1) \implies \alpha(f_L) \times \Delta u = \psi_f'(f_L).$$

Suppose there exist at least two equilibria. There are 2 possibilities: $f_L^1 = 0$, $f_L^2 > 0$; $f_L^1 = 1$, $f_L^2 < 1$ (we cannot have 2 interior equilibria).

(i) $f_L^1 = 0$, $f_L^2 \in (0, 1)$. We must have:

$$\Delta u \leq \psi_f'(0),$$

$$\Delta u > \left( 1 + \frac{1 - q}{q} f_L^2 \right) \psi_f'(0).$$

Hence, we must have

$$\left( 1 + \frac{1 - q}{q} f_L^2 \right) \psi_f'(0) < \Delta u < \psi_f'(0).$$

This cannot happen if $\psi_f'(0) = 0$ (since $\Delta u > 0$). But if $\psi_f'(0) > 0$, this is also impossible since $1 + \frac{1 - q}{q} f_L^2 > 1$.  

37
(ii) \( f_L^1 = 1, \; f_L^2 < 1. \) Then,

\[
\frac{\psi_f'(1)}{q} \leq \Delta u < \left(1 + \frac{1-q}{q} f_L^2 \right) \psi_f'(1),
\]

which is only possible if \( \frac{1}{q} \leq 1 + \frac{1-q}{q} f_L^2. \) Rearranging yields

\[
1 - q \leq (1-q) f_L^2 \iff 1 \leq f_L^2,
\]

which is again a contradiction. This establishes uniqueness.

Claim 1 follows by inspection. Let the cost of manipulation be \( \psi_f(f_L, \kappa), \) where \( \kappa \) parametrizes the marginal cost of memory manipulation: \( \frac{\partial^2 \psi}{\partial f_L \partial \kappa} > 0. \) Therefore, higher \( \kappa \)'s lead to a higher marginal cost of memory manipulation. Then, a simple computation shows that \( f_L^* \) is decreasing in \( \kappa. \) Condition (5) gives \( \frac{df_L}{dq} > 0. \) Thus, for interior solutions \( f_L^* \) is decreasing in \( q: \)

Note that the range of parameters where \( f_L^* = 0 \) is constant in \( q \) whereas the range of parameters where \( f_L^* = 1 \) is increasing in \( q. \) This establishes Claim 3.

Proof of Theorem 1:

Claims (i) and (ii) follow straight from equation (8) and the fact that

\[
\begin{align*}
q &= 0 \implies f_L^* = 0 \implies w(0) = 0 - \frac{\psi_f(0)}{\Delta u} = 0, \\
q &= 1 \implies r_H^* = 0 \implies w(1) = 1 - \frac{\psi_r(0)}{\Delta u} = 1.
\end{align*}
\]

Equation (8) implies that, \( w(q) \leq q \leq 1 \) for all \( q, \) which establishes (iii).

It remains to be shown that \( r_H^* = 0 \) or \( f_L^* = 0 \) imply that \( w(q) \geq 0. \) First, consider the case where \( f_L^* = 0. \) Then, the weighting function given by equation (8) is

\[
w(q) = q - \frac{q \psi_r(r_H^*)}{\Delta u},
\]

which is positive if and only if \( \Delta u \geq \psi_r(r_H^*). \) The argument follows by revealed preference on \( r_H^* \). Recall that it solves

\[
\max_{r_H} u_L + \{(\eta_H + r_H)[1 - \alpha(0, r_H^*)] + \alpha(0, r_H^*)\} \Delta u - \psi_r(r_H).
\]

In particular, this expression must be greater than when it is evaluated at \( r_H = 0: \)

\[
u_L + \{(\eta_H + r_H)[1 - \alpha(0, r_H^*)] + \alpha(0, r_H^*)\} \Delta u - \psi_r(0).
\]

Hence,

\[
r_H[1 - \alpha(0, r_H^*)] \Delta u \geq \psi_r(r_H) - \psi_r(0).
\]

But, since \( \alpha(0, r_H^*) \in [0, 1], r_H^* \in [0, 1], \) and \( \psi_r(0) = 0, \) we obtain \( \Delta u \geq \psi_r(r_H^*), \) which shows that \( w(q) \geq 0. \)

Next, consider the case where \( r_H^* = 0. \) Note that \( w(q) \geq 0 \) if and only if \( \frac{q}{1-q} \Delta u \geq \psi_f(f_L^*). \) The argument also follows by revealed preference. Recall that \( f_L^* \) solves

\[
\max_{f_L^*} u_L + \alpha(f_L^*, 0)(1 - \eta_L + f_L) \Delta u - \psi_f(f_L^*).
\]
Rearranging, yields

\[ u_L + \alpha (f^*_L, 0) (1 - \eta_L + f^*_L) \Delta u - \psi_f (f^*_L) \geq u_L + \alpha (f^*_L, 0) (1 - \eta_L) \Delta u - \psi_f (0). \]

Rearranging, yields

\[ \frac{q f^*_L}{q + (1 - q) f^*_L} \Delta u \geq \psi_f (f^*_L), \]

where we have used the fact that \( \alpha (f^*_L, 0) = \frac{q}{q + (1 - q) f^*_L} \). But \( \frac{q f^*_L}{q + (1 - q) f^*_L} \leq \frac{q}{1 - q} \) implies that

\[ \frac{q}{1 - q} \Delta u \geq \psi_f (f^*_L), \]

which shows that \( w (q) \geq 0. \quad \blacksquare \]

**Proof of Lemma 2:**

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

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

Since \( \Pr (\sigma = H | \theta) \) is strictly increasing in \( \theta \), 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. \( \blacksquare \)

**Proof of Lemma 3:**

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

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

**Claim 1** 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

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

39
Denote \( \# H \) for \( \# \{ t : \sigma_t = H \} \) and \( \# L \) for \( \# \{ t : \sigma_t = L \} \). Then,

\[
\begin{align*}
\frac{(\eta_H + r_t)^{\# H} (\eta_L - f_t)^{\# L} \times q^{\# H} \times (1-q)^{\# L} \times f(q)}{\prod_{t: \sigma_t = \emptyset} [q (1-\eta_H - r_t) + (1-q) (1-\eta_L + f_t)]} & \\
= \frac{\prod_{t: \sigma_t = \emptyset} [q (1-\eta_H - r_t) + (1-q) (1-\eta_L + f_t)] \times q^{\# H} \times (1-q)^{\# L} \times f(q)}{\int_{t: \sigma_t = \emptyset} [q (1-\eta_H - r_t) + (1-q) (1-\eta_L + f_t)] \times q^{\# H} \times (1-q)^{\# L} \times f(q) \, dq}.
\end{align*}
\]

Note that this is not a function of \( r_t \) and \( f_t \) for \( t \) such that \( \sigma_t \neq \emptyset \). Integrating, we obtain

\[
F(x|h^n) = \frac{\int_{0}^{x} \prod_{t: \sigma_t = \emptyset} [q (1-\eta_H - r_t) + (1-q) (1-\eta_L + f_t)] \times q^{\# H} \times (1-q)^{\# L} \times f(q) \, dq}{\int_{0}^{1} \prod_{t: \sigma_t = \emptyset} [q (1-\eta_H - r_t) + (1-q) (1-\eta_L + f_t)] \times q^{\# H} \times (1-q)^{\# L} \times f(q) \, dq}.
\]

We are now ready to prove the Claim above.

**Proof of the Claim.** We have to show that

\[
\begin{align*}
\frac{\int_{0}^{x} \prod_{t: \sigma_t = \emptyset} [q (1-\eta_H - r_t) + (1-q) (1-\eta_L + f_t)] \times q^{\# H} \times (1-q)^{\# L} \times f(q) \, dq}{\int_{0}^{1} \prod_{t: \sigma_t = \emptyset} [q (1-\eta_H - r_t) + (1-q) (1-\eta_L + f_t)] \times q^{\# H} \times (1-q)^{\# L} \times f(q) \, dq} & \\
\leq & \\
\frac{\int_{0}^{x} \prod_{t: \sigma_t = \emptyset} [q (1-\eta_H - r_t) + (1-q) (1-\eta_L + f_t)] \times q^{\# H-1} \times (1-q)^{\# L+1} \times f(q) \, dq}{\int_{0}^{1} \prod_{t: \sigma_t = \emptyset} [q (1-\eta_H - r_t) + (1-q) (1-\eta_L + f_t)] \times q^{\# H-1} \times (1-q)^{\# L+1} \times f(q) \, dq},
\end{align*}
\]

When \( x = 0 \), both sides become 0 and, when \( x = 1 \), both sides are equal to 1.
Note that \( \frac{dRHS}{dq} > \frac{dLHS}{dq} \) if and only if
\[
\prod_{t: \sigma_t = \emptyset} [x \eta_H - r_t + (1 - x) \eta_L + f_t] x^H \times (1 - x)^L \times f(x)
\]
\[
\int_0^1 \prod_{t: \sigma_t = \emptyset} [q \eta_H - r_t + (1 - q) \eta_L + f_t] q^H \times (1 - q)^L \times f(q) dq
\]
\[
\leq \prod_{t: \sigma_t = \emptyset} [x \eta_H - r_t + (1 - x) \eta_L + f_t] x^H - 1 \times (1 - x)^L+1 \times f(x)
\]
\[
\int_0^1 \prod_{t: \sigma_t = \emptyset} [q \eta_H - r_t + (1 - q) \eta_L + f_t] q^H - 1 \times (1 - q)^L+1 \times f(q) dq
\]
Rearranging, we obtain:
\[
\frac{x}{(1 - x)} \frac{\int_0^1 \prod_{t: \sigma_t = \emptyset} [q \eta_H - r_t + (1 - q) \eta_L + f_t] q^H \times (1 - q)^L \times f(q) dq}{\int_0^1 \prod_{t: \sigma_t = \emptyset} [q \eta_H - r_t + (1 - q) \eta_L + f_t] q^H - 1 \times (1 - q)^L+1 \times f(q) dq}.
\]
Thus, \( \frac{dRHS}{dq} > \frac{dLHS}{dq} \) if and only if
\[
\rho(x) > \frac{\int_0^1 \prod_{t: \sigma_t = \emptyset} [q \eta_H - r_t + (1 - q) \eta_L + f_t] q^H \times (1 - q)^L \times f(q) dq}{\int_0^1 \prod_{t: \sigma_t = \emptyset} [q \eta_H - r_t + (1 - q) \eta_L + f_t] q^H - 1 \times (1 - q)^L+1 \times f(q) dq},
\]
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
\[
\rho(x) > \frac{\int_0^1 \prod_{t: \sigma_t = \emptyset} [q \eta_H - r_t + (1 - q) \eta_L + f_t] q^H \times (1 - q)^L \times f(q) dq}{\int_0^1 \prod_{t: \sigma_t = \emptyset} [q \eta_H - r_t + (1 - q) \eta_L + f_t] q^H - 1 \times (1 - q)^L+1 \times f(q) dq} \quad \text{if } x < \bar{x},
\]
\[
\rho(x) < \frac{\int_0^1 \prod_{t: \sigma_t = \emptyset} [q \eta_H - r_t + (1 - q) \eta_L + f_t] q^H \times (1 - q)^L \times f(q) dq}{\int_0^1 \prod_{t: \sigma_t = \emptyset} [q \eta_H - r_t + (1 - q) \eta_L + f_t] q^H - 1 \times (1 - q)^L+1 \times f(q) dq} \quad \text{if } x > \bar{x}.
\]
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 3.** As shown previously, \( F(x|h^n) \) is not a function of \( r_k \) and \( f_k \) for \( k \) such that \( \sigma_k \neq \emptyset \). Therefore, we only need to establish the results for \( k \) such that \( \sigma_k = \emptyset \). Consider...
an arbitrary \( k \) such that \( \sigma_k = \emptyset \). Then, \( F(x|h^n) \) is equal to

\[
(1 - \eta_H - r_k) \int_0^x q^{#H+1} \times (1 - q)^{#L} \prod_{t \neq k; \sigma_t = \emptyset} [q (1 - \eta_H - r_t) + (1 - q) (1 - \eta_L + f_t)] f(q) dq
\]
\[
+ (1 - \eta_L + f_k) \int_0^x q^{#H} \times (1 - q)^{#L+1} \prod_{t \neq k; \sigma_t = \emptyset} [q (1 - \eta_H - r_t) + (1 - q) (1 - \eta_L + f_t)] f(q) dq
\]

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

\[
\frac{\int_0^x q^{#H} \times (1 - q)^{#L+1} \prod_{t \neq k; \sigma_t = \emptyset} [q (1 - \eta_H - r_t) + (1 - q) (1 - \eta_L + f_t)] f(q) dq}{\int_0^1 q^{#H} \times (1 - q)^{#L+1} \prod_{t \neq k; \sigma_t = \emptyset} [q (1 - \eta_H - r_t) + (1 - q) (1 - \eta_L + f_t)] f(q) dq}
\] >

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. \( \blacksquare \)

**Proof of Proposition 6:**

The result is immediate from inequality 11, Lemma 2, and the fact that the sets of histories with zero measure is the same for all relevant manipulation efforts. \( \blacksquare \)

**Proof of Proposition 7:**

In period \( N \), conditions 1 and 2 from the definition of a PBE imply that \( f_L (L, h^{N-1}) \) maximizes

\[
(\eta_L - f_L) \int u(\theta) dF(\theta|L, h^{N-1}) + (1 - \eta_L + f_L) \int u(\theta) dF(\theta|\emptyset, h^{N-1}) - \psi_f(f_L),
\]

and \( r_H (H, h^{N-1}) \) maximizes

\[
(\eta_H + r_H) \left[ \int u(\theta) dF(\theta|H, h^{N-1}) \right] + (1 - \eta_H - r_H) \left[ \int u(\theta) dF(\theta|\emptyset, h^{N-1}) \right] - \psi_r(r_H).
\]

From Proposition 6, 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 \( f_L (L, h^{N-1}) \) maximizes

\[
(\eta_L - f_L) u(\theta) + (1 - \eta_L + f_L) u(\theta) - \psi_f(f_L) = u(\theta) - \psi_f(f_L),
\]

which is strictly decreasing in \( f_L \). Hence, by continuity, it follows that \( f_L (L, h^{N-1}) \rightarrow 0 \) (a.s.). Similarly, \( r_H (H, h^{N-1}) \) maximizes \( u(\theta) - \psi_r(r_H) \), which implies that \( r_H (H, h^{N-1}) \rightarrow 0 \) (a.s.).

42
Under assumption 2b, it follows that there exists a $\Delta \bar{u}$ such that if $\int u(\theta) dF(\theta|H, h^{N-1}) - \int u(\theta) dF(\theta|L, h^{N-1}) < \Delta \bar{u}$, then $r_H(H, h^N) = f_L(L, h^{N-1}) = 0$. Because $\int u(\theta) dF(\theta|H, h^{N-1}) - \int u(\theta) dF(\theta|L, h^{N-1}) \to 0$ (a.s.), it follows that there exists an $\bar{N}$ such that $N > \bar{N}$ implies that $r_H(H, h^{N-1}) = f_L(L, h^{N-1}) = 0$ for almost all histories. The continuation payoff at $N - 1$ is then equal to $V(h^{N-1}) = u(\theta)$. Plugging back in the payoff at period $N - 1$ it follows that, for large enough $N$, $f_L(L, h^{N-2}) = r_H(H, h^{N-1}) = 0$ (a.s.). By induction, it follows that, when $N$ is large enough, there exists an $\bar{N}$ such that $r_H(h^n) = f_L(h^n) = 0$ for almost all $h^n$.

Proof of Corollary 1:

Immediate from equation (10) and Proposition 6.

Proof of Proposition 9:

The ex-ante utility net of manipulation costs is

$$\int \left\{ (1 - q) [\eta_L - f_L(t)] u_L(t) + q [\eta_H + r_H(t)] u_H(t) \right\} dH(t).$$

Substituting $u_\omega(t) = \alpha \times u_H(t) + (1 - \alpha) \times u_L(t)$ and rearranging, we obtain

$$\int \left\{ [1 - q (r_H + \eta_H)] u_L + q (r_H + \eta_H) u_H \right\} dH(t).$$

where $\alpha = \frac{q(1-r_H-\eta_H)dH(t)}{(1-q)(1-r_H-\eta_H)dH(t)+q(1-r_H-\eta_H)dH(t)}$ and we omit the dependence of $\alpha$, $u$, $r_H$, and $f_L$ on $t$ for notational simplicity.

The expected utility from a lottery whose monetary outcomes are independent of $\theta$ is $U^I = \int_t \left[ q u_H(t) + (1 - q) u_L(t) \right] dH(t)$. After some algebraic manipulations, we can express the ex-ante utility net of manipulation costs as

$$U^I + \int \left[ (g_H - g_L) \frac{q(1 - q)(1 - r_H - \eta_H)(1 - \eta_L + f_L)}{q(1 - r_H - \eta_H) g_H + (1 - q)(1 - r_L - \eta_L) g_L} \Delta u \right] dH.$$

Define the functions $\xi(t), z(t)$ as

$$z(t) = \frac{q(1 - q)(1 - r_H - \eta_H)(1 - \eta_L + f_L)}{q(1 - r_H - \eta_H) g_H + (1 - q)(1 - r_L - \eta_L) g_L} > 0,$$

$$\xi(t) = [g(t - H) - g(t - L)] z(t).$$

Then, the ex-ante utility incorporating the expected manipulation costs is equal to

$$U^I + \Lambda - MC,$$

where $\Lambda \equiv \int_t \xi(t) \Delta u(t) h(t) dt$ and $MC \equiv \int_t \left[ q \psi_r(r_H^*(t)) + (1 - q) \psi_f(f_L^*(t)) \right] h(t) dt$.

Proof of Theorem 2:

From Proposition, $U(\mathcal{L}) = q \int u_H(t) dH(t) + (1 - q) \int u_L(t) dH(t) + \Lambda - MC$. Using the definition of $v_s$, we obtain

$$U(\mathcal{L}) = q v_H + (1 - q) v_L + \Lambda - MC.$$
Using Bayes’ rule, we obtain:

\[ w(q) = q + \frac{\Lambda - MC}{v_H - v_L}. \]

Using the fact that \( v_H - v_L = \int \Delta u(t) h(t) dt \) concludes the first part of the proof.

Note also that when \( q = 0 \) or \( q = 1 \), it follows that \( MC = 0 \) and \( \Lambda = 0 \) (since \( z(t) = 0 \) for all \( t \)). Thus, \( U(L) = qv_H + (1 - q)v_L \) so that \( w(q) = q \).

**Proof of Lemma 4:**

Special case of Lemma 5.

**Proof of Lemma 5:**

We will use the following result:

**Claim 2** If \( H = L = 0 \), then \( \int \left[ qu_H(t) + (1 - q) u_L(t) \right] \left[ qg(t - H) + (1 - q) g(t - L) \right] dt = \int \int u(\theta, t) dF(\theta) g(t) dt \)

**Proof.** Recall that \( u_H(t) = \int u(\theta, t) dF(\theta) |\sigma = H \) and \( u_L(t) = \int u(\theta, t) dF(\theta) |\sigma = L \). Furthermore, \( qg(t - H) + (1 - q) g(t - L) = g(t) \). Then,

\[ \int \left[ qu_H(t) + (1 - q) u_L(t) \right] g(t) dt = \int \left[ q \int u(\theta, t) dF(\theta) |\sigma = H \right] + (1 - q) \int u(\theta, t) dF(\theta) |\sigma = L \] \( g(t) dt. \)

Using Bayes’ rule, we obtain:

\[ \int \left[ q \int u(\theta, t) dF(\theta) |\sigma = H \right] + (1 - q) \int u(\theta, t) dF(\theta) |\sigma = L \] \( g(t) dt = \int \int u(\theta, t) dF(\theta) g(t) dt. \)

The certainty equivalent is defined as

\[ \int \left\{ \left[ qu_H(t) + (1 - q) u_L(t) \right] - \left[ q\psi_H \left( r_H(t; \delta) \right) + (1 - q) \psi_L \left( f_L(t; \delta) \right) \right] + \xi(t; \delta) \Delta u(t) \right\} dt \]

\[ = \int \int u(\theta, t + CE(\delta)) dF(\theta) g(t) dt. \] (22)

Evaluating at \( \delta = 0 \) and using the result from Claim 2, we obtain:

\[ \int \left\{ - \left[ q\psi_H \left( r_H(t; 0) \right) + (1 - q) \psi_L \left( f_L(t; 0) \right) \right] + \xi(t; 0) \Delta u(t) \right\} g(t) dt \]

\[ = \int \int \left[ u(\theta, t + CE(0)) - u(\theta, t) \right] dF(\theta) g(t) dt. \]

Because \( \xi(t; 0) = 0 \), we have

\[ - \int \left[ q\psi_H \left( r_H(t; 0) \right) + (1 - q) \psi_L \left( f_L(t; 0) \right) \right] g(t) dt \]

\[ = \int \int \left[ u(\theta, t + CE(0)) - u(\theta, t) \right] dF(\theta) g(t) dt. \]

Thus:
Thus, it follows that

\[ \int f [q \psi_H (r_H (t; 0)) + (1 - q) \psi_L (f_L (t; 0))] \, g (t) \, dt = 0 \implies CE (0) = 0, \]
\[ \int f [q \psi_H (r_H (t; 0)) + (1 - q) \psi_L (f_L (t; 0))] \, g (t) \, dt > 0 \implies CE (0) < 0. \]

If \( Pr (r_H (t; 0) > 0) > 0 \) or \( Pr (f_L (t; 0) > 0) > 0 \), then \( CE (0) < 0 \) and \( \pi (0) > 0 \). Therefore, \( \lim_{\delta \to 0} \frac{\pi (\delta)}{\delta} = +\infty \) and the agent exhibits first-order risk aversion.

If \( Pr (r_H (t; 0) > 0) = Pr (f_L (t; 0) > 0) = 0 \), then \( CE (0) = 0 \) and \( \pi (0) = 0 \). Using L'Hospital's rule, it follows that

\[ \lim_{\delta \to 0} \frac{\pi (\delta)}{\delta} = \pi' (0) = -CE' (0). \]

But, differentiating equation (22) gives:

\[ CE' (0) = \int f \left\{ \left[ q \psi_H' (0) \times \frac{\partial \mu}{\partial s} (t; 0) + (1 - q) \psi_L' (0) \times \frac{\partial f}{\partial s} (t; 0) \right] \right\} g (t) \, dt \]
\[ \int f \int g \frac{du}{dt} \, (\theta, t) \, dF (\theta) \, g (t) \, dt , \]

Thus,

\[ \pi' (0) = \int f \left[ q \psi_H' (0) \times \frac{\partial \mu}{\partial s} (t; 0) + (1 - q) \psi_L' (0) \times \frac{\partial f}{\partial s} (t; 0) \right] g (t) \, dt + \phi (q) \int f g' (t) \Delta u (t) \, dt \]
\[ \int f \int g \frac{du}{dt} \, (\theta, t) \, dF (\theta) \, g (t) \, dt , \]

where \( \phi (q) = \varepsilon_1 \frac{q (1 - \eta_H) (1 - \eta_L)}{q (1 - \eta_H) + (1 - q) (1 - \eta_L)} > 0. \)

Because the denominator is positive, it follows that:

- \( \pi' (0) = 0 \) if \( \int f \left[ q \psi_H' (0) \times \frac{\partial \mu}{\partial s} (t; 0) + (1 - q) \psi_L' (0) \times \frac{\partial f}{\partial s} (t; 0) \right] g (t) \, dt + \phi (q) \int f g' (t) \Delta u (t) \, dt = 0, \)
- \( \pi' (0) > 0 \) if \( \int f \left[ q \psi_H' (0) \times \frac{\partial \mu}{\partial s} (t; 0) + (1 - q) \psi_L' (0) \times \frac{\partial f}{\partial s} (t; 0) \right] g (t) \, dt + \phi (q) \int f g' (t) \Delta u (t) \, dt > 0, \) (first-order risk-averse) and
- \( \pi' (0) < 0 \) if \( \int f \left[ q \psi_H' (0) \times \frac{\partial \mu}{\partial s} (t; 0) + (1 - q) \psi_L' (0) \times \frac{\partial f}{\partial s} (t; 0) \right] g (t) \, dt + \phi (q) \int f g' (t) \Delta u (t) \, dt < 0 \) (first-order risk-loving),

which concludes the proof.

Proof of Corollary 2:

Recall that \( u (\theta, t) = \phi (\theta) \times t \) and \( \int \phi (\theta) f (\theta) \, d\theta = 1 \). Thus,

\[ \int_{t_0} \int_{\theta} u (\theta, t_0 + L) \, f (\theta) \, g (t_0) \, d\theta \, dt_0 = L + E (t_0). \]

Moreover, \( \int_{\theta} \int_{t_0} u (\theta, t) \, f (\theta) \, h (t) \, d\theta \, dt = \int t \times h (t) \, dt = E (t) \). Since \( t = t_0 + s \), it follows that

\[ E (t) = E (t_0) + qH + (1 - q) L. \]

Substituting in equation (16), we obtain

\[ MC \leq q (H - L) + \Lambda. \]
which concludes the proof.  ■

Proof of Proposition 12:

From Proposition 8, it follows that \(MC > 0\) for any \(t \geq 0\), which establishes the first claim. Moreover, for \(\Delta t = 0\), we have that \(MC \leq q\Delta t\) in any equilibrium. By the theory of the maximum, it follows that \(MC \leq q\Delta t\) when \(\Delta t > 0\) is sufficiently small.  ■

Proof of Lemma 7:

Let \(\phi\) denote \(E(\phi|\text{Termination})\). Terminating a project after observing \(t^N\) gives an expected payoff of

\[
[(\eta_N + r_N - f_N) \phi_N + (1 - \eta_N - r_N + f_N) \tilde{\phi}] \times E(t_0) - MC_N,
\]

whereas continuing gives \(\phi_N \times [t^N + E(t_0)]\). Thus, the project is continued if

\[
\phi_N t^N + (1 - \eta_N - r_N + f_N) E(t_0) (\phi_N - \tilde{\phi}) \geq -MC_N,
\]

which is true because the left-hand side is positive whereas the right-hand side is negative. Thus, the project is not terminated.

Hence, in any PBE where \(\text{Termination} \neq \emptyset\), Bayes’ rule implies that \(\tilde{\phi} \leq \phi_{N-1}\). The result then follows by induction.  ■

Proof of Proposition 13:

Suppose that both inefficient projects 1 and 2 are terminated. Then, equation (19) yields

\[
-\phi_i t^i \geq (1 - \eta_i - r_i + f_i) \left( \phi_i - \frac{q_1 \phi_1 + q_2 \phi_2}{q_1 + q_2} \right) \times E(t_0) + MC_i,
\]

for \(i \in \{1, 2\}\). Rearranging yields, for \(i \in \{1, 2\}\),

\[
t^i \leq -(1 - \eta_i - r_i + f_i) \left( 1 - \frac{1}{\phi_i} \frac{q_1 \phi_1 + q_2 \phi_2}{q_1 + q_2} \right) \times E(t_0) - \frac{MC_i}{\phi_i}.
\]

Under Assumption 2a, \(\inf \{MC_i : t \geq 0\} > 0\). Therefore, whenever \(t^i > -\frac{\inf \{MC_i : t \geq 0\}}{\phi_i}\) for some \(i \in \{1, 2\}\), project \(i\) is not terminated. Thus, setting \(\tilde{t}_i = -\frac{\inf \{MC_i : t \geq 0\}}{\phi_i} < 0\) establishes that terminating 1 and 2 cannot be a PBE.  ■

B  Example

In this Section, I present an example of explicit distributions of \(\theta\) and \(\sigma\). Suppose \(\theta \sim U[0, 1]\) and

\[
\sigma = \begin{cases} H & \text{if } \theta \geq 1 - q, \\ L & \text{if } \theta < 1 - q. \end{cases}
\]

Then, \(F(\theta|\sigma = H) = \frac{\theta}{q} \times \chi(\theta \geq 1 - q)\) and \(F(\theta|\sigma = L) = \frac{\theta}{1 - q} \times \chi(\theta < 1 - q)\), where \(\chi\) denotes the indicator function. It follows that

\[
\begin{align*}
u_H &= \frac{1}{q} \int_{1-q}^1 u(\theta) \, d\theta, \quad \text{and} \\
u_L &= \frac{1}{1-q} \int_0^{1-q} u(\theta) \, d\theta.
\end{align*}
\]
Hence, \( \Delta u(q) = \frac{1}{\sqrt{2}} \int_{1-q}^{1} u(\theta) d\theta - \frac{1}{\sqrt{q}} \int_{0}^{1-q} u(\theta) d\theta \) which, as stated in Subsection 2.4, is in general a function of \( q \). \( \Delta u(q) \) can be an increasing or decreasing function of \( q \) depending on the utility function \( u(\theta) \).

For example, when \( u(\theta) = \theta^{2} \), then \( \Delta u(q) = \frac{2}{3} - \frac{1}{3}q \), which is decreasing in \( q \). When \( u(\theta) = \theta \), then \( \Delta u(q) = \frac{1}{2} \). And, when \( u(\theta) = \sqrt{\theta} \), then \( \Delta u(q) = \frac{2}{3q} (1 - \sqrt{1-q}) \), which is increasing in \( q \).

C Non-Bayesian Framework

In this Section, I consider deviations from the assumption that the decision-maker evaluates her expected characteristics when a signal is forgotten according to Bayes' rule. Suppose, instead, that upon recollecting \( \hat{\sigma} = \emptyset \), the DM attributes weight

\[
\rho(f_{L}, r_{H}) = \frac{q(1 - \eta_{H} - r_{H})}{q(1 - \eta_{H} - r_{H}) + \iota(1 - q)(1 - \eta_{L} + f_{L})}
\]

to \( \sigma = H \) and \( 1 - \rho(f_{L}, r_{H}) \) to \( \sigma = L \), where \( \iota \in [0, 1] \) denotes the degree of pessimism in the agent’s updating rule. When \( \iota = 0 \), the agent believes that any forgotten signal is a high signal. Thus, she updates information optimistically. When \( \iota = +\infty \), the agent believes that forgotten signals are low signals and, therefore, updates information pessimistically. Finally, \( \iota = 1 \) captures the Bayesian case described in the text (realistic updating).

It is straightforward to show that the ex-ante expected utility from observing the signal is equal to

\[
1 - (1 - q)(\eta_{L} - f_{L}) + (1 - q)(\eta_{L} - f_{L})(\eta_{H} + r_{H})(1 - \iota) - (\eta_{H} + r_{H}) + \iota(1 - q)(\eta_{H} + r_{H})q\Delta u
\]

\[
+ u_{L} - q\psi_{H}(r_{H}) - (1 - q)\psi_{L}(f_{L}).
\]

For \( \iota = 0 \), the expression above becomes \( u_{H} - (1 - q)(\eta_{L} - f_{L})\Delta u - q\psi_{H}(r_{H}) - (1 - q)\psi_{L}(f_{L}) \). In that case, the DM would always choose \( r_{H}^{*} = 0 \) (since there is no loss from forgetting a high signal when \( \hat{\sigma} = \emptyset \) are interpreted as \( \sigma = H \)). Thus, the ex-ante expected utility is

\[
u_{H} - (1 - q)(\eta_{L} - f_{L}^{*})\Delta u - (1 - q)\psi_{L}(f_{L}^{*})
\]

The agent prefers to observe the signal if and only if

\[
(1 - \eta_{L} + f_{L}^{*})\Delta u \geq \psi_{L}(f_{L}^{*})
\]

Hence, if the manipulation costs are not too high, the DM may prefer to observe a signal when she has biased recollections in the sense of being more optimistic than implies by Bayes’ rule.

When \( \iota = +\infty \), the ex-ante expected utility from observing the signal is

\[
u_{L} + q(\eta_{H} + r_{H})\Delta u - q\psi_{H}(r_{H}) - (1 - q)\psi_{L}(f_{L})
\]

Since this expression is always smaller than \( qu_{H} + (1 - q)u_{L} \), the DM will never prefer to observe the signal when \( \iota = +\infty \). Indeed, this result is true for any agent with recollections that are more pessimistic than implied by Bayes’ rule (\( \iota > 1 \)).

Extending this approach to the case of lotteries over money, it can be shown that deviations from Bayes’ rule introduce an additional term in the representation of Proposition 9. In that
case, the DM may prefer a lottery whose outcomes are correlated with characteristics even if there are no complementarities between characteristics and money.

It is interesting to note that the results from Section 3, where it was shown that the DM’s behavior converges to the one predicted by expected utility theory after observing a sufficiently large number of signals, do not necessarily require the agent to update according to Bayes rule or even to eventually learn her true type $\theta$.

Consider, for example, the case of extreme optimism: $\varepsilon = 0$. In this case, the agent interprets $\hat{\sigma} = \emptyset$ as $\sigma = H$. Hence, even though she does not update recollections according to Bayes’ rule, the agent’s inference problem can be reinterpreted as if she observed signals $\sigma = L$ with probability $\left(1-q\right)\left(\eta_L - f_L\right)$ and $\sigma = H$ with probability $1 - \left(1-q\right)\left(\eta_L - f_L\right)$ and applied Bayes’ rule.

Following the same arguments as in Section 3, we can show that the Bayes estimator $\hat{u}_n$ converges to some function $\tilde{u}(\theta)$ for almost every history when $N \to +\infty$. In general, we would have $\tilde{u}(\theta) \neq u(\theta)$ so that an extremely optimistic agent would not eventually learn her true type. Nevertheless, because $\hat{u}_n$ converges, the benefit of memory manipulation converges to zero. Therefore, the agent’s behavior also converges to the one of an expected utility maximizer despite the fact that she does not update according to Bayes rule and does not eventually learn her true type $\theta$.

References


