# Online Appendix

## A Existence of Equilibrium

**Proposition 12** *There exists a PBE.*

**Proof:** Fix an ex-ante action $a \in A$, and consider the "restricted game" $G_a$ as the continuation game following such action. From Definition 1, a profile of PBEs of all restricted games $\{\phi_a^*, \phi_b^*, m_H^*(a), m_L^*(a), \mu(.|\hat{\sigma}_a)\}_{a\in A}$ is a PBE of the original game considered in the text if and only if

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

First, we note that the all continuation games have a PBE. Fix an action $a \in A$ (which, for notational simplicity, will be omitted from the expressions below). Let $b_H \in \arg\max_b u_H(b,H)$, $b_L \in \arg\max_b u_L(b,L)$, and note that both $b_H$ and $b_L$ are not functions of $m_H, m_L$, or $\alpha$. Define the correspondence $T : [-\eta_H, 1-\eta_H] \times [-\eta_L, 1-\eta_L] \times [0,1] \times \Delta(B) \to [-\eta_H, 1-\eta_H] \times [-\eta_L, 1-\eta_L] \times [0,1] \times \Delta(B)$ by

$$T\left(m_H, m_L, \alpha, \phi_{b_\varnothing}\right) \equiv \left(m_H^*\left(m_L, \phi_{b_\varnothing}, \alpha\right), m_L^*\left(m_H, \phi_{b_\varnothing}, \alpha\right), \alpha\left(m_L, m_H, \phi_{b_\varnothing}\right), \phi_{b_\varnothing}\left(m_L, m_H, \alpha\right)\right),$$

where:

$$m_H^* \equiv \arg\max_{m_H} (\eta_H + m_H) u_H(b_H, H) + (1-\eta_H - m_H)\left[\alpha^* u_H\left(\phi_{b_\varnothing}, H\right) + (1-\alpha^*) u_L\left(\phi_{b_\varnothing}, H\right)\right] - \psi_H(m_H), \quad (23)$$

$$m_L^* \equiv \arg\max_{m_L} (\eta_L + m_L) u_L(b_L, L) + (1-\eta_L - m_L)\left[\alpha^* u_H\left(\phi_{b_\varnothing}, L\right) + (1-\alpha^*) u_L\left(\phi_{b_\varnothing}, L\right)\right] - \psi_L(m_L), \quad (24)$$

$$\alpha^* = \frac{q(1-\eta_H - m_H^*)}{q(1-\eta_H - m_H^*) + (1-q)(1-\eta_L - m_L^*)}, \quad (25)$$

$$\phi_{b_\varnothing} \equiv \arg\max_{\{\phi(b)\}_{b\in B}} \left\{ \sum_b \phi(b)\left[\alpha^* u_H(b,H) + (1-\alpha^*) u_L(b,L)\right] \right\}. \quad (26)$$

A fixed point of $T$, combined with $b^*(H) = b_H$ and $b^*(L) = b_L$ constitute a PBE of the restricted game $G_a$. Equation (23) is a continuous and concave function of $m_H$. Therefore, $m_H^*$ is a non-empty, convex, and compact set. Similarly, equation (24) is a continuous and concave function of $m_L$, implying that $m_L^*$ is a non-empty, convex, and compact set. Note that equation (25) defines $\alpha^*$ as a function of $m_L, m_H, b_\varnothing$. Therefore, $\alpha^*$ is a singleton, which is trivially a non-empty, convex, and compact set. Moreover, since the equation in (26) is a linear function of $\phi(b)$, $\phi_{b_\varnothing}$ is a non-empty, convex, and compact set. Then, Kakutani's theorem establishes that there exists a fixed-point of $T$, which corresponds to a PBE of the restricted game $G_a$.

Given a profile of PBEs of all restricted games $\{G_a : a \in A\}$, define

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

Since $A$ is finite, such an $a^*$ exists. The profile $\{a^*, \phi_b^*, m_H^*(a), m_L^*(a), \mu(.|\hat{\sigma}_a)\}$ constructed previously constitutes a PBE of the original game. ■

## B Non-Bayesian Framework

Throughout the paper, I have maintained the assumption that decision makers understand that they engage in memory manipulation and, thus, interpret their recollections according to Bayes' rule (sophisticated decision makers). In this section, I consider the case of naive individuals. As in Mullainathan (2002), naive individuals are unaware of their imperfect memory and interpret recollections as if they were the true outcomes.

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

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

$$\rho \equiv \frac{q(1-\eta_H)}{q(1-\eta_H)+(1-q)(1-\eta_L)} \tag{27}$$

to a high outcome and $(1-\rho)$ to a low outcome. I refer to this rule as a *naive updating rule.*[60] The following definition proposes an adaptation of the PBE concept to naive decision makers:

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

1. $a^* \in \arg\max_{a \in A} \left\{ \begin{array}{c} E_{\hat{\sigma}_a}\left[E_\mu\left[u\left(a, b_a^*(\hat{\sigma}_a), \theta, \sigma_a\right)|\hat{\sigma}_a\right]|m_L^*(a), m_H^*(a)\right] \\ -q\psi_H\left(m_H^*(a)\right) - (1-q)\psi_L\left(m_L^*(a)\right) \end{array} \right\}$;
2. $m_\sigma^*(a) \in \arg\max_{m_\sigma} \left\{ \begin{array}{c} (\eta_\sigma + m_\sigma) E_\mu\left[u\left(a, b_a^*(\hat{\sigma}_a), \theta, s\right)|\hat{\sigma}_a = \sigma\right] \\ +(1-\eta_\sigma - m_\sigma) E_\mu\left[u\left(a, b_a^*(\hat{\sigma}_a), \theta, \sigma\right)|\hat{\sigma}_a = \varnothing\right] - \psi_\sigma(m_\sigma) \end{array} \right\}$,
   $\sigma \in \{H, L\}$;
3. $b_a^*(\hat{\sigma}) \in \arg\max_{b \in B} \left\{E_\mu\left[u(a, b, \theta, \sigma_a)|\hat{\sigma}_a = \hat{\sigma}\right]\right\}$;
4. $\mu(\theta|\hat{\sigma}_a = \hat{\sigma})$ *is obtained by the naive updating rule if* $\Pr(\hat{\sigma}_a = \hat{\sigma}|m_L = m_H = 0, a) > 0$, $\hat{\sigma} \in \{L, H, \varnothing\}$, $a \in A$.

Conditions $1-3$ are the same as in the PBE concept. Condition 4 modifies the standard Bayesian condition by requiring agents to follow the naive updating rule instead. An immediate consequence of the naive updating rule in this model is that it leads to posterior beliefs that first-order stochastically dominate beliefs obtained by Bayes' rule. Therefore, naive individuals are optimistic about their type $\theta$.

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

## B1 Equilibrium Uniqueness

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

$$u_\varnothing(a, b, \sigma) = \rho u_H(a, b, \sigma) + (1-\rho) u_L(a, b, \sigma). \tag{28}$$

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

$$(\eta_s + m_s) u_s\left(a, b_a(s), s\right) + (1-\eta_s - m_s) u_\varnothing\left(a, b_a(\varnothing), s\right) - \psi_s(m_s). \tag{29}$$

The key feature of the naive updating rule is that it is not a function of the amount of memory manipulation employed by self 1. This greatly simplifies the computation of the PNE of the model since, unlike in the sophisticated case, there is no feedback between self 2's expectation of the manipulation exerted by self 1 and self 1's manipulation choice. Then, the equilibrium amount of manipulation is determined by the maximum of expression (29).

[60] *Benabou and Tirole (2006a) present a slightly different formulation of naiveté. In their model, agents follow an updating rule that underweights bad news by a fixed proportion. Unlike under our notion of naiveté, they still obtain multiple equilibria when agents are naive. However, as in our model, naiveté also tends to reduce the welfare costs of self-deception.*

**Proposition 13** *There exists a PNE. Furthermore, if $\psi_\sigma$ is strictly convex and $b_a(\hat{\sigma})$ is a (single-valued) function where $\sigma \in \{H, L\}$ and $\hat{\sigma} \in \{H, L, \varnothing\}$, the PNE is essentially unique.*[61]

**Proof.** Existence of PNE follows an argument analogous to the existence of a PBE. Note that Condition 3 from Definition 3 implies that $b_a(\hat{s})$ is not a function of self 1's memory manipulation. Strict convexity of $\psi_s$ implies that expression (29) is strictly concave in $m_s$. Then, the equilibrium amounts of memory manipulation $m_L^*$ and $m_H^*$ are unique. Condition 4 implies that beliefs must also coincide in all recollections such that $\Pr(\hat{\sigma}_a = \hat{\sigma} | m_L^* = m_H^* = 0) > 0$. ■

**Corollary 2** *The PNE is essentially unique when either: (i) $u(a, ., \theta, \sigma_a) : B \to \mathbb{R}$ is a strictly concave function, or (ii) $B$ is a singleton (i.e., the individual does not take ex-post actions).*

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

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

## B2 Ambiguity-Seeking Behavior

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

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

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

**Remark 8** *Proceeding as in Appendix A, the NDM's expected utility from the monetary lottery can be represented by*

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

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

## B3 Zeroth-Order Risk Seeking Behavior

This subsection shows that the NDM may be zeroth-order risk seeking. As in Subsection A.3, consider a lottery that pays $x = \varepsilon s$, $s \in \{H, L\}$, where $qH + (1-q)L = 0$. Let $m_s^*(\varepsilon)$ denote the equilibrium amount of memory manipulation as a function of $\varepsilon$. The certainty equivalent of this lottery is:

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

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

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

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

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

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

**Proposition 14** *The NDM is:*

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

## B4 Convergence

This subsection shows that, under certain regularity conditions, the beliefs of a NDM converge to the truth despite her departure from Bayes' rule. Let $\tilde{\theta}_n$ denote the NDM's expected type.

**Proposition 15** *Let $N \to \infty$ and fix a NPE. There exists a random variable $z$ such that $\tilde{\theta}_n \to z$ for almost all histories.*

**Proof.** Let $\tilde{q}_n$ denote the expected value of $q$ according to the naive updating rule. It is straightforward to verify that $1 - \tilde{q}_n$ follows a supermartingale and, by Doob's theorem, it converges in distribution. Then using the fact that $q(\theta)$ is continuous and strictly increasing concludes the proof. ■

The proposition above does not imply that manipulation converges to zero since $z$ may be nondegenerate and, therefore, $\Delta u$ may not converge to zero. In order to ensure that this is the case, we will use the following regularity condition:

**Regularity** Condition. The prior distribution $f(q)$ satisfies

$$\lim_{n\to\infty} \left\{ \begin{array}{c} \frac{\int_0^x q^{\alpha n+1}(1-q)^{\beta n}[q(1-\eta_H)+(1-q)(1-\eta_L)]^{(1-\alpha-\beta)n} f(q)dq}{\int_0^1 q^{\alpha n+1}(1-q)^{\beta n}[q(1-\eta_H)+(1-q)(1-\eta_L)]^{(1-\alpha-\beta)n} f(q)dq} \\ -\frac{\int_0^x q^{\alpha n}(1-q)^{\beta n+1}[q(1-\eta_H)+(1-q)(1-\eta_L)]^{(1-\alpha-\beta)n} f(q)dq}{\int_0^1 q^{\alpha n}(1-q)^{\beta n+1}[q(1-\eta_H)+(1-q)(1-\eta_L)]^{(1-\alpha-\beta)n} f(q)dq} \end{array} \right\} = 0,$$

for all $(\alpha, \beta, x) \in [0,1]^3$.

The regularity condition states that the impact of a single observation is small when the number of observations is large. It is satisfied, for example, in the forgetfulness model of Example 1 and in the limited memory model of Example 2 when the prior distribution is uniform. The following proposition shows that although the NDM is optimistic in any finite period, her beliefs converge to the truth and her memory manipulation converges to zero.

**Proposition 16** *Suppose the regularity condition is satisfied and let $N \to \infty$. Then, $m_\sigma \xrightarrow{D} 0$ and $\hat{\theta}_n \xrightarrow{D} \theta$.*

**Proof.** Let $\hat{q}(h^n)$ denote the NDM's beliefs about $q$ given $h_n$. The regularity condition implies that $(\hat{q}_t|h^n, H) - (\hat{q}_t|h^n, L) \xrightarrow{D} 0$. Let $\tilde{E}$ be the expectation under the NDM's beliefs. Then, $\tilde{E}[q|h^n, H] - \tilde{E}[q|h^n, L] \to 0$. Then, the amount of memory manipulation converges to zero. But $m_s \xrightarrow{D} 0$ implies that $\hat{\theta}_n \xrightarrow{D} \theta$. ■

# C Non-Separable Preferences

In Section A, the DM's preferences additively separable between attributes and money. In this appendix, I consider general utility functions. A key ingredient of the general model is the degree of complementarity between attributes and money. Since a DM is not as affected by monetary outcomes when she is uninformed about her attributes when attributes and money are complementary, complementarity can be interpreted as providing "psychological insurance." Therefore, the DM may prefer a lottery whose outcomes are informative about her attributes if the complementarity effect is greater than the costs of self-deception. Moreover, the resulting probability weighting function may have an "inverted S-shape" as in Tversky and Kahneman (1992) and Prelec (1998).

Let $u_\sigma(s) \equiv \int u(\theta, s)\, dF(\theta|\sigma)$ denote the expected utility from a monetary amount equal to $s$ conditional on outcome $\sigma$. From Lemma 2, $m_H^* > 0 \geq m_L^*$. Define the degree of complementarity between $\theta$ and money by

$$\chi(H, L) \equiv u_H(H) + u_L(L) - u_L(H) - u_H(L). \tag{30}$$

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

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

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

$$U(\Sigma) = qu_H(H) + (1-q) u_L(L) + z\chi(H, L) - MC. \tag{31}$$

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

The utility of a monetary lottery can be decomposed in three terms: First, the expected utility $qu_H(H) + (1-q) u_L(L)$ of the lottery when memory is perfect. Second, the expected manipulation costs $MC$. These two effects are precisely the same as in the additively separable case (see equation 11). The third effect, which is not present when the utility is additively separable, is the degree of complementarity between attributes and money. When signals are forgotten, there is probability $\alpha(m_H^*, m_L^*)$ that a high signal was observed and the complementary probability that a low signal was observed. Thus, forgetting a signal can be seen as providing "psychological insurance" to the agent. This raises her expected utility if $\theta$ and money are complementary ($\chi > 0$) and decreases her expected utility if they are substitutes ($\chi < 0$).

Proceeding as in Subsection A.1, the DM's expected utility can be represented by

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

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

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

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

As in Section A, denote by $U^I$ the utility of a lottery with the same distribution over monetary outcomes as the one above but whose monetary outcomes are uninformative about $\theta$. Rearranging equation (31), we obtain

$$U(\Sigma) = U^I + y\chi(H, L) - MC, \tag{32}$$

where $y = q(1-q)\left[1 + \frac{(1-\eta_L - m_L^*)(1-\eta_H - m_H^*)}{q(1-\eta_H - m_H^*) + (1-q)(1-\eta_L - m_L^*)}\right] > 0$. Consider the choice between the lottery $\Sigma$ and another lottery with the same distribution over monetary outcomes but whose monetary outcomes are uninformative about $\theta$. Equation (32) implies that the DM will prefer lottery $\Sigma$ if the degree of complementarity is high enough or if the expected memory cost is low enough: $y\chi(H, L) \geq MC$. Therefore, when attributes and money are complementary, the DM may prefer the attribute-dependent lottery.

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

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

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

**Proposition 17** *In any PBE, the DM exhibits zeroth-order risk aversion.*

**Proof.** For any PBE, define the expected manipulation cost as $MC(\varepsilon) \equiv q\psi_H(m_H^*(\varepsilon)) + (1-q)\psi_H(m_L^*(\varepsilon))$. Note that $\lim_{\varepsilon\to 0_+}\chi(\varepsilon H, \varepsilon L) = 0$. Therefore, for small $\varepsilon$, equation (15) becomes:

$$\int u(\theta, CE(\varepsilon))\, dF(\theta) = qu_H(\varepsilon H) + (1-q)\, u_L(\varepsilon L) - MC(\varepsilon). \tag{33}$$

Since $MC(0) > 0$ and, by the Theorem of the Maximum, $MC(\varepsilon)$ is continuous, it follows that $MC(\varepsilon) > 0$ for small $\varepsilon$. Hence, $\lim_{\varepsilon\to 0_+} MC(\varepsilon) > 0$. Then, equation (33) yields:

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

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

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

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

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

**Proposition 18** *In the exogenous memory model: (i) the DM is first-order risk averse if $u_L'(0) > u_H'(0)$; (ii) the DM is first-order risk seeking if $u_L'(0) < u_H'(0)$; and (iii) the DM has second-order risk preferences if $u_L'(0) = u_H'(0)$.*

**Proof.** Since $MC(\varepsilon) = 0$ for all $\varepsilon$, equation (15) becomes

$$\int u(\theta, CE(0))\, dF(\theta) = qu_H(\varepsilon H) + (1-q)\, u_L(\varepsilon L) + z\chi(\varepsilon H, \varepsilon L). \tag{34}$$

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

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

Therefore, Bayes' rule implies that $\int u(\theta, CE(0))\, dF(\theta) = \int u(\theta, 0)\, dF(\theta)$ and, because $u$ is strictly increasing in money, $\pi(0) = -CE(0) = 0$. Differentiating equation (34), it follows that

$$CE'(0) = \frac{qu_H'(0)\, H + (1-q)\, u_L'(0)\, L + z(H-L)[u_H'(0) - u_L'(0)]}{qu_H'(CE) + (1-q)\, u_L'(CE)}.$$

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

$$CE'(0) = K[u_H'(0) - u_L'(0)],$$

where $K = \frac{H}{qu_H'(0) + (1-q)u_L'(0)}\left(q + \frac{z}{1-q}\right) > 0$. Then, by L'Hospital's rule, it follows that

$$\lim_{\varepsilon\to 0_+}\pi(\varepsilon)/\varepsilon = -CE'(0) = -K[u_H'(0) - u_L'(0)],$$

which concludes the proof. ■

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

## D Proof of Claim 2

The p.d.f. of $q$ conditional on $h^n$ is

$$f\left(q|h^n\right)=\frac{\begin{array}{c}\prod\limits_{t:\sigma_t=\varnothing}\left[q\left(1-\eta_H-m_{H,t}^*\right)+\left(1-q\right)\left(1-\eta_L-m_{L,t}^*\right)\right]\\ \times\prod\limits_{t:\sigma_t=H}q\left(\eta_H+m_{H,t}^*\right)\times\prod\limits_{t:\sigma_t=L}\left(1-q\right)\left(\eta_L+m_{L,t}^*\right)\times f\left(q\right)\end{array}}{\int\left\{\begin{array}{c}\prod\limits_{t:\sigma_t=\varnothing}\left[q\left(1-\eta_H-m_{H,t}^*\right)+\left(1-q\right)\left(1-\eta_L-m_{L,t}^*\right)\right]\\ \times\prod\limits_{t:\sigma_t=H}q\left(\eta_H+m_{H,t}^*\right)\times\prod\limits_{t:\sigma_t=L}\left(1-q\right)\left(\eta_L+m_{L,t}^*\right)\times f\left(q\right)\end{array}\right\}dq}.$$

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

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

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

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

We are now ready to prove the Claim:

**Proof of Claim 2.**We have to show that

$$\begin{array}{c}\frac{\int_0^x\prod\limits_{t:\sigma_t=\varnothing}\left[q\left(1-\eta_H-m_{H,t}^*\right)+\left(1-q\right)\left(1-\eta_L-m_{L,t}^*\right)\right]\times q^{\#H}\times\left(1-q\right)^{\#L}\times f\left(q\right)dq}{\int_0^1\prod\limits_{t:\sigma_t=\varnothing}\left[q\left(1-\eta_H-m_{H,t}^*\right)+\left(1-q\right)\left(1-\eta_L-m_{L,t}^*\right)\right]\times q^{\#H}\times\left(1-q\right)^{\#L}\times f\left(q\right)dq}\leq\\ \frac{\int_0^x\prod\limits_{t:\sigma_t=\varnothing}\left[q\left(1-\eta_H-m_{H,t}^*\right)+\left(1-q\right)\left(1-\eta_L-m_{L,t}^*\right)\right]\times q^{\#H-1}\times\left(1-q\right)^{\#L+1}\times f\left(q\right)dq}{\int_0^1\prod\limits_{t:\sigma_t=\varnothing}\left[q\left(1-\eta_H-m_{H,t}^*\right)+\left(1-q\right)\left(1-\eta_L-m_{L,t}^*\right)\right]\times q^{\#H-1}\times\left(1-q\right)^{\#L+1}\times f\left(q\right)dq}.\end{array}$$

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

The derivative of the LHS with respect to $x$ is

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

and the derivative of the RHS with respect to $x$ is

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

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

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

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

Rearranging, we obtain:

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

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

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

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

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

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

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

# References


Prelec, D. (1998). "The probability weighting function," Econometrica, 66, 497–527.

Tversky, A. and D. Kahneman (1992). "Advances in prospect theory: cumulative representation of uncertainty," Journal of Risk and Uncertainty, 5, 297–323.