Selective Memory and Motivated Delusion: Theory and Experiment

Soo Hong Chew[*] Wei Huang[†] Xiaojian Zhao[‡]

February 2013

Abstract

Building on the works of Carrillo and Mariotti (2000) and Benabou and Tirole (2002), we formulate an intra-person, multiple-self model of how motivated memory including amnesia and delusion may relate to the individual's degree of present bias. We posit the notion of (non-)conscious choice in which the individual habituates into being forgetful or delusional to enhance the motivation for one's future selves. In equilibrium, the model endogenizes the individual's state of motivated memory and particularly delusion resulting from a high level of present bias. We test our model in a controlled incentivized experiment using the Ravens IQ test and find overall support for its implications except for the significant incidence of positive confabulation. This leads us to extend our basic model to capture this possibility as an equilibrium outcome.

Keywords: selective memory, motivated delusion, self control, intrapersonal game
JEL Classification: D03, D83, Z13

[*] Department of Economics and Department of Finance, National University of Singapore, Singapore. Email: chew.soohong@gmail.com

[†] Department of Economics, Hong Kong University of Science and Technology, Hong Kong. E-mail: whuang@ust.hk

[‡] Department of Economics, Hong Kong University of Science and Technology, Hong Kong. E-mail: xjzhao@ust.hk

> ‘Consider, Sir,’ answered Sancho, ‘that those which appear yonder, are not giants, but windmills; and what seem to be arms are the sails, which, whirled about by the wind, make the millstone go.’
>
> ‘One may easily see,’ answered Don Quixote, ‘that you are not versed in the business of adventures: they are giants; and, if you are afraid, get aside and pray, whilst I engage with them in a fierce and unequal combat.’

Miguel de Cervantes, *Don Quixote*, (1605, Chapter 8).

1 Introduction

Implicit in much of economic analysis is the assumption of full consciousness – the decision maker has unlimited ability for attention towards stimuli, unbounded capacity for recording and storage of events, and perfect accuracy of recall. This runs counter to the fact that as biological beings, we have limited capacity for attention, recording, and recall. At any moment, we cannot be conscious of all stimuli or sensations registered at the biological level nor can we recall accurately all that are stored at various levels of memory. To varying degrees, people are susceptible to limited attention and imperfect recall including being possibly delusional (see, e.g., Pashler, 1998).

A growing number of papers in the literature contribute to filling this gap by studying various aspects of bounded consciousness on economic decision making.[1] Among them, Carrillo and Mariotti (2000) study rational inattention from the perspective of strategic ignorance – whether to not acquire information costlessly prior to its resolution – in an intra-person, multiple-self setting. Adopting a similar setting, Benabou and Tirole (2002) study selective memory in terms of one’s decision on whether to delete information after it is received. Both models rely on the incidence of present bias. In the Carrillo-Mariotti model of positive denial, present bias induces a need to sustain personal motivation by ignoring information that may weaken the individual’s self confidence. In the Benabou-Tirole model of positive amnesia, present bias induces a corresponding tendency to forget signals believed to be negative.[2]

[1] Since Simon (1947), economists have attempted models in which individuals simplify complex decisions by processing only a subset of information. More recently, Dow (1991) studies optimal search under limited memory and shows how a decision maker may focus scarce cognitive resources on part of the problem. On a distinct note, Gottlieb (2010) makes use of imperfect memory and self-deception to obtain a non-expected utility representation to account for a range of observed anomalies. Building on the idea of rational inattention, Sims’ (2003) model of sources of inertia in prices and wages has inspired several follow up papers.

[2] Besides the notion of personal over confidence, Svenson (1981) studies over confidence in terms of social rankings. In a recent experimental test, Brown, Croson and Eckel (2011) find support of Carrillo-Mariotti’s model. Burks et al (2012) design an experiment to test three mechanisms that

People also tend to forget negative experiences more readily than positive ones.[3] When they exhibit false memory, people tend to be delusional in the positive direction. In an extreme case, Ramachandran (1996) documents how a women who could not move her left arm claims that she could engage in activities that require the use of both hands, say clapping. In principle, anosognosia (not being conscious of one's disability) enables one to preserve a positive self-image in the face of potentially debilitating adversities.[4] More recently, McKay and Dennett (2009), Howe and Derbish (2010) and Howe et. al. (2011) show how delusion can have positive consequences in conjunction with the adaptive function of memory and deliver fitness-relevant benefits for subsequent behavior and problem solving.

The present paper extends the Benabou-Tirole model of selective recall by incorporating the possibility of positive delusion, i.e., remembering a good signal when there was none, in addition to selective recall. In the Benabou-Tirole model, the incidence of present bias provides a channel for the individual to form motivated beliefs and exhibit selective memory in equilibrium in an intra-person, multi-self setting. Our paper expands the possible mental states beyond the set of recalled experiences to include the possibility that individuals may nonconsciously create certain fake signals.

Our model endogenizes the individual's state of motivated memory including delusion and relates it to his attitude towards intertemporal discounting. Following Benabou and Tirole (2002), present bias is a primitive in our model leading the individual to undertake less investment activity than otherwise. The individual arrives at his belief nonconsciously inducing a demand for over confidence to motivate himself to resolve this under-investment problem in terms of ex ante evaluation. On the supply side, besides positive amnesia, positive delusion may offer an additional channel to deliver over confidence. Otherwise, having a correct belief about oneself can lead to under-invest due to present bias. In our model, delusion and amnesia can act as substitute motivational mechanisms for the individual. Unlike Benabou and Tirole (2002), for sufficiently severe present bias, the possibility of delusion precludes amnesia from delivering over confidence. Consequently, the incidence of positive amnesia

may deliver over confidence: Bayesian updating (Benoit and Dubra, 2011), concern for self image (Koszegi, 2006; Weinberg, 2006) and social signaling (Benabou and Tirole, 2002). They reject the first two mechanisms and argue that overconfidence is a social signaling bias.

[3] See the review papers in psychology (Dunning, 2001; Knudsen, 2007) and a recent experimental study in economics (Li, 2012).

[4] Another example is the reverse Othello syndrome involving a delusory in the fidelity of one's partner, which can works as a defensive mechanism to maintain one's self-esteem (see Burler (2000) and McKay, Langdon, and Coltheart (2005) for a more detailed discussion). Relatedly, Bortolotti and Mameli (2012) argue that delusions of persecutions explain away the fact of one's failure, and attribute his failure entirely to someone's conspiracy against him. Thus motivational factors also contribute to some delusions of persecutions.

does not depend monotonically on the magnitude of present bias.

We conduct an experiment to test the implications of our model in two stages. In the initial stage, subjects take an incentivized Ravens IQ test after completing a number of decision making tasks including one on temporal discounting. In a subsequent stage months later, subjects are shown 6 questions. Four are from the original test while two are similar but new. The appearance of each question is accompanied the correct answer. Subjects are asked to recall whether they (a) did it correctly, (b) did it incorrectly, (c) did not see it, or (d) do not remember. Subjects receive S$1 for each correct recall, lose S$1 for each incorrect recall, and receive nothing for "I do not remember". The experiment, involving 768 subjects in Singapore, generates data on memory pattern, degree of present bias, and risk attitude besides demographic information.

While the implications of our model are generally supported by our experimental findings, we observe a significant incidence of positive confabulation (false memory in which a bad signal is transformed into a good signal) which runs counter to our basic model. We consider two distinct mechanisms to realize confabulation: (1) a bad signal is transformed into a good signal directly in one step; (2) a bad signal is omitted in the first step (amnesia) followed by the creation of a good signal subsequently (delusion). We note that in the latter mechanism, confabulation necessitates delusion as a necessary route from a bad signal to a good one. This is corroborated by the finding in our experiment that individuals exhibiting confabulation also tend to exhibit positive delusion, and that individuals without delusion do not tend to exhibit confabulation. This two-step mechanism appears also compatible with the idea of Korsakoff Syndrome (Whitty and Lewin, 1960) in which confabulation may serve a compensatory pseudo-reminiscence role to fill memory gaps. In other words, the brain can produce false memory to make up for memory loss.

Motivated by the finding of pervasiveness of confabulation in our experiment, we extend our basic model by incorporating the possibility of positive confabulation (from bad to good signal) in addition to positive amnesia (from bad to no signal) and positive delusion (from none to good signal). The observed concurrence of positive delusion with positive confabulation further suggests that the process of confabulation may involve two steps – forgetting a bad signal followed by fantasizing about having received a good signal. This leads us to extend our basic model by allowing for the possibility of positive confabulation in two steps rather than directly transforming a bad signal into a good signal. The equilibrium behavior of the extended model, encompassing positive amnesia, positive delusion, and positive confabulation, is largely supported by our experimental findings.

The paper is organized as follows. Section 2 presents our basic model. Section 3 discusses our experimental design and findings. Section 4 presents our extended model with confabulation. Section 5 discusses the implications of our model and experimental findings in three subsections – collectivist interpretations, consciousness, and mental well being. We conclude in Section 6.

2 Basic Model

Building on Carrillo and Mariotti (2000) and Benabou and Tirole (2002), the approach adopted here is game-theoretic and deliberately stylized to study intra-person interactions involving multiple selves. The specific model described here extends the Benabou-Tirole model of selective memory by incorporating additionally the possibility of delusion.

To capture the divergence in interests at different epochs between possibly distinct selves, we assume three epochs, $t = 0, 1, 2$. Memory is imperfectly formed during $t = 0$ and contributes to the individual's belief at $t = 1$ when decisions are made. At the outset, the individual (self-0) may receive a private signal s concerning his ability θ, e.g., achievements, feedback from work, and social interactions. We denote by s $= B$ for a bad signal, $s = G$ for a good signal, and $s = \emptyset$, for no signal. As the individual approaches $t = 1$, he may recall correctly or incorrectly whether he did receive a signal. We consider two kinds of recall errors: (i) forget seeing a signal, and (ii) recall seeing a signal when there was none. During $t = 1$, the individual (self-1) engages in an activity at cost c which if successful will yield benefit V at $t = 2$ and zero otherwise. Self-0 knows the distribution of c which we assume for simplicity to be uniformly distributed over $[0, \bar{c}]$. The individual's ability is captured by the probability of success θ in the costly investment activity.

Following the hyperbolic discounting literature (see, e.g., Strotz, 1955; Laibson, 1997), self-t discounts the expected utility of self$-(t + n)$ at $\beta\delta^n$ for $n = 1, 2, 3, \cdots$ where $\delta = 1$ is the normal discount rate and $\beta < 1$ corresponds to the incidence of present bias. We view the individual as a collection of incarnations with possibly divergent goals.[5]

We now describe how we specifically model selective memory and delusion. Consider the case where $s = \emptyset$ (no signal) with probability $1 - q$, $s = B$ (bad signal) with probability qp, and $s = G$ (good signal) with probability $q(1 - p)$. Here, q can be interpreted as a measure of the individual's degree of social exposure from which he receives feedback from social interactions. Thus, conditional on receiving a signal, the probability of it being bad is given by p and being good is given by $1 - p$. Given

[5] We may interpret the different epochs as stages in an atemporal multi-stage setting to model selective attention in decision making and allow for the possibility of fantasies. The game may also be recast in an inter-generational setting at the societal level to study collective amnesia and myth making as we will discuss in Section 5.

feedback from society, p represents the chance that the individual will face a negative event.

For $s = B$, $\emptyset$, or G, we refer to the individual as type-B, type-$\emptyset$, or type-G, respectively. Let θ_s refer to the expected value of θ conditional on each possible realization of the true signal s, i.e., $\theta_B = E[\theta|s = B]$, $\theta_\emptyset = E[\theta|s = \emptyset]$ and $\theta_G = E[\theta|s = G]$. Let

$$\theta_\emptyset = p\theta_B + (1 - p)\theta_G.$$

where $\theta_B \ < \theta_G$. That is, receiving no signal implies that his ability is given by the expected ability in the presence of a signal.

As in Benabou and Tirole (2002), our model posits "present" bias as a primitive which induces a demand by the individual for over confidence. Besides suppressing the bad signal for type-B self, we incorporate the additional possibility of creating a good signal for type-$\emptyset$ self.

Here s is the objective signal the individual obtains from the external world. In order to incorporate self-0's role in memory formation including the possibility of delusion for self-1, we let $\hat{s}$ denote the subjective signal transmitted from self-0 to self-1. Specifically:

(i) $s = G$: In this case $\hat{s} = G$ in the absence of opportunity for signal manipulation.

(ii) $s = B$: In this case $\hat{s} = B$ or $\hat{s} = \emptyset$. Self-0 may communicate the signal truthfully to self-1 or suppress the bad signal (*amnesia*).

(iii) $s = \emptyset$: In this case $\hat{s} = \emptyset$ or $\hat{s} = G$. Self-0 may leave it as it is or create a fake good signal (*delusion*).

As discussed in the Introduction, there is a tendency for selective memory and delusion to exhibit a positive bias which we capture here by our present setup.[6] One may further consider the possibility of confabulation, i.e., transforming a bad signal $s = B$ into a good signal one $\hat{s} = G$. We will address this issue in a more elaborate model in Section 4 following our experimental results.

Let $h_s = \Pr[\hat{s} = s|s]$ denotes the probability that he chooses to transmits the signal s truthfully to self-1 by self-0 of type s. We denote by h_B^* and $h_\emptyset^*$ the respective beliefs held by self-1 concerning self-0 being truthful in the chosen recall and delusion strategies.

The intra-personal game involving the individual's memory strategy is depicted in Figure 1.

People often forget and sometimes fantasize. Casual introspection suggests that human being cannot consciously choose to forget or to be delusional. Yet, as dis-

[6] While we can accommodate the additional possibilities of forgetting a good signal for type-G self as well as creating a bad signal for type-$\emptyset$ self under a "future" bias inducing a demand for under confidence, doing so would introduce undue complexities to the model without enhancing its ability to account for positive bias in recall and delusion.

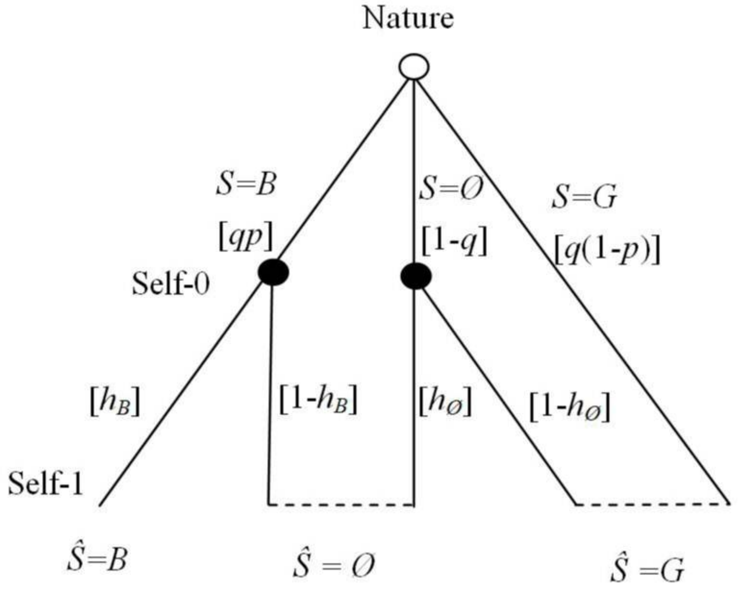


Figure 1: Memory Strategy

cussed in the Introduction, while memory bias including delusion is inherently not a conscious process, they are known to possess directionality in terms of a tendency to forget bad signals as well as to fantasize positively. Since the seminal work of Hebb (1949), ample evidence in neuroscience shows that environmental stimuli give rise to dramatic changes in brain functions including altered learning and memory process (see, e.g., Williams et al., 2001; Meshi et al., 2006). As such, we posit that self-0 memory choices, both amnesia and delusion, are made nonconsciously. Our modeling assumption allowing for forgetting bad signals and romanticizing fake signals reflects the empirical finding as reported in McKay and Dennett (2009), Howe and Derbish (2010) and Howe et. al. (2011). Notably, in a review paper, Howe (2011) suggests that delusional dispositions can have functionality in terms of adaptation. Nonetheless, we do observe people engaging consciously in specific acts to facilitate their forgetting certain bad signals, e.g., leaving a place to avoid bad memories or burning photos of ex-spouses, and also to induce fake but good signals, such as reading chivalric novels such as “Don Quixote” or idling about in “Second Life” in the internet age, possibly to enhance one’s subsequent self image.

At epoch 1, self-1 forms expectations over his ability θ in light of the recalled $\widehat{s}$, taking into account the possibility that self-0 may have suppressed the true signal or created a fake signal. Let $\theta^*(\widehat{s})$ denote self-1’s assessment of his ability given $\widehat{s}$ and $r^*(\widehat{s})$ denote the reliability of the signal $\widehat{s}$, i.e., the probability that the signal $\widehat{s}$ is accurate. When $\theta^*(\widehat{s}) > \theta_s$, we say the individual exhibits *over confidence*. Clearly, $\theta^*(B) = \theta_B$. For the case of $\hat{s} = \emptyset$, applying Bayes’ rule, we have:

$$r^*(\emptyset) = \Pr[s = \emptyset | \hat{s} = \emptyset; h_B^*; h_\emptyset^*] = \frac{(1-q)\,h_\emptyset^*}{qp(1-h_B^*) + (1-q)\,h_\emptyset^*}.$$

It follows that

$$\theta^*(\emptyset) = r^*(\emptyset)\theta_\emptyset + (1 - r^*(\emptyset))\theta_B \geq \theta_B.$$

Similarly, for the case of $\hat{s} = G$, we have:

$$r^*(G) = \Pr[s = G | \hat{s} = G; h_B^*; h_\emptyset^*] = \frac{q(1-p)}{(1-q)\,(1-h_\emptyset^*) + q(1-p)},$$

and

$$\theta^*(G) = r^*(G)\theta_G + (1 - r^*(G))\theta_\emptyset > \theta_\emptyset.$$

The last inequality reflects strict over confidence of self-1 when $s = \emptyset$, because his updated belief $\theta^*(G)$ of his ability is always higher than the true ability $\theta_\emptyset$ of self-0 of type-$\emptyset$ when self-1 receives a good signal. However, his updated ability given no signal $\theta^*(\emptyset)$ equals θ_B when $h_\emptyset^*$ equals 0 so that $r^*(\emptyset) = 0$. In other words, the incidence of delusion precludes amnesia from delivering over confidence.

Self-1 will incur the cost of investment if and only if

$$\beta\theta^* V - c \geqslant 0.$$

The divergence in interest between self-0 and self-1 is captured by having $\beta < 1$. Notice that the qualitative features of the analysis above is robust with respect to a form of non-Bayesianism often called partial naivete in which the first term of the denominator in the expression for reliability is modified by a factor λ which may be less than 1.

When $s = B$, self-0 chooses recall strategy h_B. Should he transmit the signal accurately to self-1 ($\hat{s} = B$), his expected utility would be given by:

$$U_{CR}(\theta_B) = \int_0^{\beta\theta_B V} \{\theta_B V - c\} dF(c)$$

where F refers to the distribution function of c (not necessarily uniformly distributed) and the subscript CR stands for "correct recall".[7] By contrast, if self-0 suppresses the bad signal ($\hat{s} = \emptyset$), his expected utility is given by:

$$U_A(\theta_B) = \int_0^{\beta\theta^*(\emptyset) V} \{\theta_B V - c\} dF(c)$$

[7] For simplicity, we ignore the present bias of self-0 when calculating his expected utility.

where the subscript A stands for "amnesia". The net gain from suppressing the bad signal is equal to:

$$U_A(\theta_B) - U_{CR}(\theta_B) = \int_{\beta\theta_B V}^{\beta\theta^*(\emptyset)V} \{\theta_B V - c\} dF(c). \tag{1}$$

When $\beta\theta^*(\emptyset)V$ exceeds $\beta\theta_B V$, amnesia delivers over confidence which in turn gives rise to more investment activities.

Similarly, when $s = \emptyset$, self-0's expected utility from transmitting the signal accurately to self-1 ($\hat{s} = \emptyset$) is given by:

$$U_{CR}(\theta_\emptyset) = \int_0^{\beta\theta^*(\emptyset)V} \{\theta_\emptyset V - c\} dF(c).$$

If on the other hand self-0 fabricates a fake signal ($\hat{s} = G$), his expected utility would be given by:

$$U_D(\theta_\emptyset) = \int_0^{\beta\theta^*(G)V} \{\theta_\emptyset V - c\} dF(c)$$

where the subscript D stands for "delusion". The net gain from creating a fake signal is then given by:

$$U_D(\theta_\emptyset) - U_{CR}(\theta_\emptyset) = \int_{\beta\theta^*(\emptyset)V}^{\beta\theta^*(G)V} \{\theta_\emptyset V - c\} dF(c). \tag{2}$$

Notice that $\beta\theta^*(G)V$ always exceeds $\beta\theta^*(\emptyset)V$, so that delusion delivers over confidence leading further to more investment activities.

2.1 Memory Bias in Equilibrium

> "In madness equilibrium is established, but it masks that equilibrium beneath the cloud of illusion, beneath feigned disorder; the rigor of the architecture is concealed beneath the cunning arrangement of these disordered violences."
>
> Foucault (1961), p.34

A central question of our work is on the relation between the individual's magnitude of present bias and his possible states of motivated memory in equilibrium. As mentioned previously, the problem of time inconsistency arises from the incidence of present bias inducing the individual to under invest in period 1. Here, over-confidence can alleviate the under-investment problem. In the basic model, amnesia and delusion provide two mechanisms to deliver over-confidence. The role of amnesia has been studied in Benabou and Tirole (2002). In our model, we explore the role of delusion and study how these motivational mechanisms contribute to the supply of over confidence, how they depend on present bias β, and how motivated delusion interacts with amnesia.

Here, we apply the solution concept of perfect Bayesian equilibrium in conjunction with the intuitive criterion refinement (Cho and Kreps, 1987) to shed light on pure-strategy equilibrium outcomes which can be tested experimentally.[8]

We begin with a simplifying restriction of no delusion, $h_{\emptyset} = 1$. This gives rise to Benabou and Tirole's (2002) model which serves as a benchmark for our basic model in this section and its extension in a subsequent section. As in Benabou and Tirole (2002), we can obtain the following proposition.

Proposition 1 *(Benabou and Tirole, 2002) The likelihood of amnesia increases in the degree of present bias.*

Here, amnesia can serve to deliver over confidence which can in turn alleviate the under-investment problem when present bias is severe. Thus the likelihood of amnesia depends negatively on β. As mentioned in the Introduction, this proposition is not supported by our experimental finding discussed in next section.

By allowing for the possibility of delusion of self-0 of type-$\emptyset$, we obtain the following existence result under the assumption of uniformly distributed cost of investment:

Proposition 2 *There are three possible perfect Bayesian equilibria in pure strategies.*

(i) (PBE1: Correct Recall) There exists perfect Bayesian equilibrium, $h_B^* = 1$, $h_{\emptyset}^* = 1$ *if* $\beta > \beta_1$ *for some* $\beta_1 \in (0, 1)$.

(ii) (PBE2: Positive Amnesia) For p *close to* 1, *there exists perfect Bayesian equilibrium,* $h_B^* = 0$, $h_{\emptyset}^* = 1$ *if* $\beta \in (\underline{\beta}_2, \overline{\beta}_2)$ *for some* $\underline{\beta}_2$, $\overline{\beta}_2 \in (0, 1)$.

(iii) (PBE3: Positive Delusion) There exists perfect Bayesian equilibrium, $h_{\emptyset}^* = 0$ *and* $h_B^* = 1$ *or* 0, *if* $\beta < \beta_3$ *for some* $\beta_3 \in (0, 1]$.

In case (i) with β large enough, i.e., the problem of present bias is not too severe, there is an equilibrium with perfect recall and no delusion. In this case, information is always valuable because of an alignment in interests between self-0 and self-1. Here, truthful reporting is an equilibrium regardless of the signal received by self-0. Otherwise, amnesia or delusion would create a bias towards over-investment of self-1 from the perspective of self-0 because of his (almost) perfect dynamic consistency. Since PBE1 does not involve amnesia or delusion, we call this equilibrium "correct recall".

[8] Under uniform distribution, we can solve the mixed strategy equilibiria explicitly. For simplicity, we do not consider mixed strategy equilibria in the balance of the paper.

In case (ii) where p is close to 1, there is an equilibrium (PBE2) with amnesia and no delusion for intermediate values of β. The amnesia condition requires β to be bounded from above while no-delusion requires β to be bounded from below. Here, self-1 of type-B and type-$\emptyset$ are pooled. If the individual receives a bad signal, he would choose to ignore it leading him to become over confident. On the other hand, if self-1 does not receive any signal, he becomes under confident since he does not know whether it is indeed the case. With p close to 1, $h_\emptyset$ is closer to h_B than h_G. In this case, a small degree of over confidence will make amnesia beneficial. At the same time, with the large gap between $h_\emptyset$ and h_G, any incidence of delusion would induce excessive over investment. Thus, PBE2 with amnesia without delusion emerges.

The equilibrium outcome in case (iii) involves two possibilities.

(a) Delusion without amnesia: Here, self-0 of type-B transmits the bad signal to self-1 while self-0 of type-$\emptyset$ creates a fake good signal. In this case, self-1 of type-$\emptyset$ and type-G are pooled. Thus self-1 of type-G will have self-doubt because he is uncertain whether what he receives is a true signal or a fake signal created by self-0, i.e., a delusion strategy delivers under confidence for self-1 of type-G. Here, delusion requires a low β while no amnesia requires a high β. It follows that an intermediate level of β is required to maintain delusion without amnesia. Note that this equilibrium is determined by the off-equilibrium belief $r^*(\emptyset)$. When $r^*(\emptyset) = 0$, self-0 of type-B is indifferent between correct recall and amnesia. When $r^*(\emptyset) > 0$, he will choose to recall the bad signal for β sufficiently large. Under the intuitive criterion, $r^*(\emptyset)$ can only be zero (see Appendix A.1 for details). In this case, the decision maker is indifferent between correct recall and amnesia and would not need a large β to remain self truthful. Thus, when β is small enough, this equilibrium prevails.

(b) Delusion with amnesia: When present bias is sufficiently severe, i.e., β small enough, self-0 of both type-B and type-$\emptyset$ will cheat. Nonetheless, this equilibrium outcome is similar to the case above. There is no actual amnesia in equilibrium because when self-1 receives no signal, he knows that he is of type-B. While delusion and amnesia can act as substitutes as motivational mechanisms, delusion precludes amnesia from delivering over confidence in this equilibrium. Compared to case (a) above, we call the amnesia in case (b) fake amnesia. Thus, the equilibria in case (a) and case (b) are essentially identical.

Corollary 1 *The likelihood of delusion increases in the degree of present bias.*

The implication of Corollary 1 is supported by Result 6 in our experimental finding discussed in the next section. A higher chance of delusion results from a greater present bias in time preference. However, in contrast to Benabou and Tirole (2002), we find that the existence of the equilibrium with amnesia does not depend monotonically on the magnitude of present bias due to the possibility of fake amnesia in case (iiib). In other words, given the existence of delusion, amnesia no loner has a role in delivering over confidence when present bias is severe. These theoretical implications are also supported by our experiment results as will be discussed in Section 3.

Proposition 3 *If PBE2 exists, then* $\overline{\beta}_2 > \beta_1$ *and* $\underline{\beta}_2 < \beta_3$.

Should PBE2 exist, the set of parameters $(\beta_1, 1)$ for the existence of PBE1 intersects with $(\underline{\beta}_2, \overline{\beta}_2)$ for the existence of PBE2. Furthermore, $(\underline{\beta}_2, \overline{\beta}_2)$ intersects with the corresponding set of parameters $(0, \beta_3)$ for PBE3. Notice that the existence of PBE2 requires p to be sufficiently large. Suppressing the bad signal is beneficial when the probability of a bad signal p is high, which makes self-1 of type-$\emptyset$ less over confident thereby reducing the risk of over-investment for self-0 of type-B. Meanwhile, a high value of p lowers the true ability of self-0 of type-$\emptyset$ compared to θ_G. Thus, having delusion is costly due to the problem of over-investment.

In general, self-1's belief and self-0's choice are complementary. Information manipulation may entail over confidence of self-0 in which case the cost is lower if self-1 is more skeptical in regarding the signal as being less reliable. Meanwhile, when self-1 exhibits more self doubt, self-0 is more likely to manipulate the signal since the cost of over investment as a result of over confidence is lower. Similarly, for self-0, being more self truthful may enhance self-1's trust in him, thereby inducing self-0 to be in turn more truthful. This suggests the existence of multiple equilibria leading to the following propositions.

Proposition 4 *If PBE2 does not exist, then* $\beta_1 < \beta_3$.

Combining the propositions above we have the following corollary on the existence of equilibria.

Corollary 2 *For any degree of present bias* β*, there exists at least one equilibrium.*

The following proposition concerns the possibility of multiple equilibria.

Proposition 5 *If PBE2 exists, for small enough* p*, we have* $\beta_1 < \beta_3$*, i.e., there exists a triplet of pure-strategy equilibria.*

Given the multiplicity of equilibria, we apply an individual-level ex ante expected utility welfare criterion to facilitate equilibrium selection. The ex ante welfare $W(PBE1)$ of the self-truthful individual is given by

$$qp\int_0^{\beta\theta_B V}(\theta_B V - c)dF(c) + (1-q)\int_0^{\beta\theta_\emptyset V}(\theta_\emptyset V - c)dF(c) + q(1-p)\int_0^{\beta\theta_G V}(\theta_G V - c)dF(c).$$

The ex ante welfare $W(PBE2)$ of the self-deceitful who is forgetful is given by

$$qp\int_0^{\beta(\frac{1-q}{qp+1-q}\theta_\emptyset+\frac{qp}{qp+1-q}\theta_B)V}(\theta_B V-c)dF(c)+(1-q)\int_0^{\beta(\frac{1-q}{qp+1-q}\theta_\emptyset\frac{qp}{qp+1-q}\theta_B)V}(\theta_\emptyset V-c)dF(c)$$
$$+\,q(1-p)\int_0^{\beta\theta_G V}(\theta_G V-c)dF(c).$$

The ex ante welfare $W(PBE3)$ of the self-deceitful who is delusional is given by

$$qp\int_0^{\beta\theta_B V}(\theta_B V-c)dF(c)+(1-q)\int_0^{\beta[\frac{q-qp}{1-qp}\theta_G+\frac{1-q}{1-qp}\theta_\emptyset]V}(\theta_\emptyset V-c)dF(c)$$
$$+\,q(1-p)\int_0^{\beta[\frac{q-qp}{1-qp}\theta_G+\frac{1-q}{1-qp}\theta_\emptyset]V}(\theta_G V-c)dF(c).$$

The following proposition shows that being self-truthful is ex ante better off than being self-deceitful regardless of whether the individual is forgetful or delusional.

Proposition 6 $W(PBE1)>\max\{W(PBE2),\, W(PBE3)\}$.

Intra-person signal manipulations will reduce one's ex ante well being. To see this, take PBE2 as an example. Here, self-0 of type-B chooses to suppress the bad signal. However, over confidence of self-1 of type-B entails under confidence of self-1 of type-$\emptyset$ (when he receives no signal), which exacerbates the problem of under-investment of self-1 of type-$\emptyset$ due to his time inconsistent preference. Amnesia motivates self-1 of type-B to undertake more activities but discourages self-1 of type-$\emptyset$ to undertake less. At the same time, the relatively more able self-1 of type-$\emptyset$ should engage in more activities. Thus, the overall ex ante welfare is lower in PBE2 than in PBE1. Similarly, in PBE3, delusion entails self-doubt of self-1 of type-G, which ultimately reduces one's ex ante welfare. Interestingly, a perfect commitment of truth-telling can enhance the ex ante welfare. However, this intra-personal commitment is not renegotiation-proof. Upon receiving a bad signal or no signal, the individual may find it advantageous to renegotiate with himself, and thereby engage in self-deception.

Thus it is preferable for the individual to be self-truthful from an ex ante perspective. If this cannot be achieved, the question whether it is better to be forgetful, i.e., $W(PBE2)$, or delusional, i.e., $W(PBE3)$ remains. The following proposition shows that $W(PBE2)>W(PBE3)$ when p is high enough.

Proposition 7 $W(PBE2)>W(PBE3)$ *if and only if* $1/2<p<1$.

When the chance of a negative event is high, it is better for the individual to be forgetful than delusional. Intuitively, in PBE2, the over-confident self-1 of type-B is the beneficiary compared to PBE3 in which self-1 of type-$\emptyset$ is the beneficiary. When p is sufficiently high, self-1 is more likely to have received $s=B$ so that PBE2 is better than PBE3.

2.2 Comparative Statics

To obtain comparative statics results for different equilibria, it is convenient to derive explicitly the critical values of β in Proposition 2 displayed in Appendix A. It is then straightforward to obtain the comparative statics of the critical values of β with respect to $\gamma = \theta_G/\theta_B$ (> 1) which we interpret to be a measure of the informativeness of the signals for an individual's ability. A higher γ indicates a more informative signal since a good signal entails a relatively higher ability compared to the estimated ability upon receiving a bad signal.

Proposition 8 *β_1, $\overline{\beta}_2$, $\underline{\beta}_2$, and β_3 are all decreasing in γ.*

Proposition 8 tells us that the incidence of each such kind of self-deceiving equilibria would require a greater degree of present bias when the signal is more informative. Intuitively, with highly informative signals on one's ability, having memory bias could lead to significant over confidence and subsequent over investment, which may be functional for highly impatient individuals. For example, in PBE1, the behavior of an individual with a more informative signal will more likely resemble a self-truthful equilibrium without amnesia or delusion. The reason is that when θ_G becomes large relative to θ_B, the smallest β that guarantees the existence of PBE1 turns out to be lower. This arises from observing that when γ increases, the upper bound of the integral in (1) (and (2)) increases at a higher rate than its lower bound. Thus β needs to be lower in order to maintain indifference. Analogous reasoning applies for PBE2 and PBE3.

We further obtain comparative statics of the critical values of β with respect to q as proxy for the individual's frequency of social interaction. The individual will receive more signals from society with higher values of q.

Proposition 9 *(i) β_1 is independent of q.*
(ii) Both $\overline{\beta}_2$ and $\underline{\beta}_2$ are increasing in q.
(iii) β_3 is decreasing in q.

Part (i) tells us that the incidence of self-truthfulness is not dependent on the frequency of social interaction since all signals are reliable. The revised belief about his ability is independent of the probability of not receiving any signal so that q does not influence β_1.

Part (ii) says that a higher frequency of social interaction requires greater patience in delivering an equilibrium with amnesia without delusion. As the individual assimilates into society, the degree of present bias would be mitigated to maintain PBE2. When q increases, the upper bound of equation (1) decreases as it gives a higher weight to θ_B without changing the lower bound. Hence, for a higher q, $\overline{\beta}_2$ needs to be larger to maintain indifference. Similarly, $\underline{\beta}_2$ increases in q.

PBE3 tells us that the incidence of delusion is reduced with a higher frequency of social interaction. Thus, solitary individuals are more likely to suffer from delusion. Intuitively, when q gets smaller, the upper bound of equation (2) will decrease. This yields a higher weight to $\theta_\emptyset$ without changing the lower bound, so that β_3 needs to increase in order to maintain indifference.

Lastly, we derive the comparative statics for the critical values of β with respect to the conditional probability of receiving a negative signal p.

Proposition 10 *(i) β_1 decreases in p as long as $p < (1 - \gamma^{-1/2})/(1 - \gamma^{-1})$, and increases thereafter.*

(ii) $\underline{\beta}_2$ decreases in p and $\overline{\beta}_2$ increases in p.

(iii) β_3 decreases in p when p is small and increases in p when p is large.

Part (i) describes how an individual who is not stressed by frequent negative signals (p small) will become more truthful with an increase in p. Yet, in the face of frequent negative signals (p large), the individual will likely become more self deceptive. Parts (ii) and (iii) tell us that a higher level of negative signals will lead to a higher chance of amnesia but a lower chance of delusion. While a dose of harsh reality can awaken one from a pipe dream, the individual retreats into amnesia rather than face reality. Intuitively, as the probability of receiving a bad signal increases, self-B benefits more from amnesia, thereby increasing the likelihood of PBE2. In (iii), when p is small enough, the individual is more likely to be type-G. In this case, a decrease in p makes $\theta_\emptyset$ to be closer to θ_G, reducing the cost of excessive over investment from delusion, and increasing the cost of excessive over investment from amnesia. Thus PBE3 is more sustainable.

We summarize the welfare comparative statics below.

Proposition 11 *For all perfect Bayesian equilibria, we have that ex ante welfare is increasing in β, increasing in q, and decreasing in p.*

Notice that present bias, being a primitive in our model, drives the divergence in interests between self-0 and self-1. Consequently, regardless of the individual's state of equilibrium, he is ex ante better off with a lower degree of present bias. Similarly, in each kind of equilibrium, it is better for the individual to have a higher chance of receiving signals. Not surprisingly, each type of individuals will be ex ante worse off with an increased probability of receiving a bad signal.

3 Experiment on Memory Bias

We report the findings of an incentivized experiment using 768 subjects recruited from the National University of Singapore to test the implications from our model linking temporal discounting and memory error proneness. Subjects' degree of present bias are elicited in terms of a comparison between their tradeoffs in a near term task (next day versus 30 days later) versus a remote term task (351 days versus 381 days later). Their patterns of memory errors are studied in a 2-stage experiment involving performance on an incentivized Ravens IQ test followed by a test of their recall (see Appendix B) of their performance months later.[9] The discounting task together with a number of risk taking tasks are administered as part of the initial stage.

The subsequent stage tests subjects' ability to recall their performance in the Ravens IQ test by showing them 6 test questions one at a time together with the correct answers. Of the 6 questions, four appeared in the initial stage and 2 are new. For each of these 6 questions, subjects can choose one of 4 responses:

- a : My response was correct.
- b : My response was incorrect.
- c : I didn't see this question.
- d : I don't remember.

All 6 questions in the second stage are incentivized: For the 4 questions which had appeared previously, subjects each receives S$1 (about US$0.80) if their choice reflects correctly their performance or if they choose (c) for the 2 questions which had not appeared previously. The subject loses S$1 when his/her choices reveals a memory error – exhibiting false memory about having seen a question which had not appeared previously or remembering incorrectly. Subjects always receive nothing if they choose (d) – "I don't remember".

For each of the 6 questions presented in the second stage, the subject either did it right ($s = G$), did it wrong ($s = B$), or did not see it ($s = \varnothing$) at the initial stage. The table below displays subjects' possible responses in relation to s.

	a	b	c	d
$s = G$	a_G : CR	b_G : Negative C	c_G : Negative A	d_G : Weak Negative A
$s = B$	a_B : Positive C	b_B : CR	c_B : Positive A	d_B : Weak Positive A
$s = \varnothing$	$a_\varnothing$: Positive D	$b_\varnothing$: Negative D	$c_\varnothing$: CR	$d_\varnothing$: Weak CR

[9] Because the tests are online, the duration between two stages are different from months to a year, which depend on the response time of the subjects.

There are three types of correct recall (CR), a_G, b_B, and $c_\varnothing$. Moreover, compared to $c_\varnothing$ (recalling correctly that one has not seen the question previously), $d_\varnothing$ (choosing "I don't remember" when one has not seen the question before) reveals a weak sense of correct recall. There remains 8 types of memory errors: two linked to delusion (D), $a_\varnothing$ and $b_\varnothing$, and two linked to confabulation (C), a_B and b_G. In terms of amnesia (A), when the question had appeared before, stating "I don't remember" (option (d)) is weaker than claiming "I didn't see this question" (option (c)). Thus, we denote d_G and d_B as weak amnesia compared with c_G and c_B as amnesia. Given the parallel between responses in (c) and the responses in (d), we have grouped them together in our data analysis.

In terms of our basic model, we classify the recalled signal $\widehat{s}$ as follows: $\widehat{s} = G$ if we observe a_G or $a_\varnothing$; $\widehat{s} = B$ if we observe b_B; and $\widehat{s} = \varnothing$ if we observe $c_\varnothing$, c_B, $d_\varnothing$, or d_B. Our model cannot account for positive confabulation – a_B – or negative recall – b_G, $b_\varnothing$, c_G, and d_G. In terms of observed behavior to be exposited shortly, the incidence of negative recall behavior is not significant. On the other hand, positive confabulation turns out to be significantly observed. This has motivated us to develop an extension of our basic model addressed in the next section.

3.1 Pattern of Memory Bias

Here, we discuss the observed patterns of our experiment on memory error proneness. The overall memory patterns (see Table C1 in Appendix C) reveals significant incidence of positive amnesia (36.8%), positive delusion (59.2%), and positive confabulation (48.9%). In each case, we find a consistent tendency for a positive bias which we shall detail below.

Result 1 *Individuals tend to exhibit positive amnesia rather than negative amnesia.*

Figure 2a displays the respective rates of positive amnesia $(c_B+d_B)/(b_B+c_B+d_B)$ for Q1 to Q4 at 83.6%, 82.3% 76.8% and 61.9%, which are significantly higher than the corresponding rates of negative amnesia $(c_G + d_G)/(a_G + c_G + d_G)$ at 34.6%, 37.9%, 40.7% and 38.3% (see Table C1). In other words, individuals who did a question incorrectly are significantly more likely to forget than those who did the question correctly. This behavior is compatible with the implications of the Benabou-Tirole model as well as our basic model.

Result 2 *Individuals tend to exhibit positive delusion rather than negative delusion.*

Figure 2b displays the rate of positive delusion $a_\varnothing/(a_\varnothing+b_\varnothing)$ and negative delusion $b_\varnothing/(a_\varnothing + b_\varnothing)$. The rates of positive delusion for Question 5 and Question 6 are

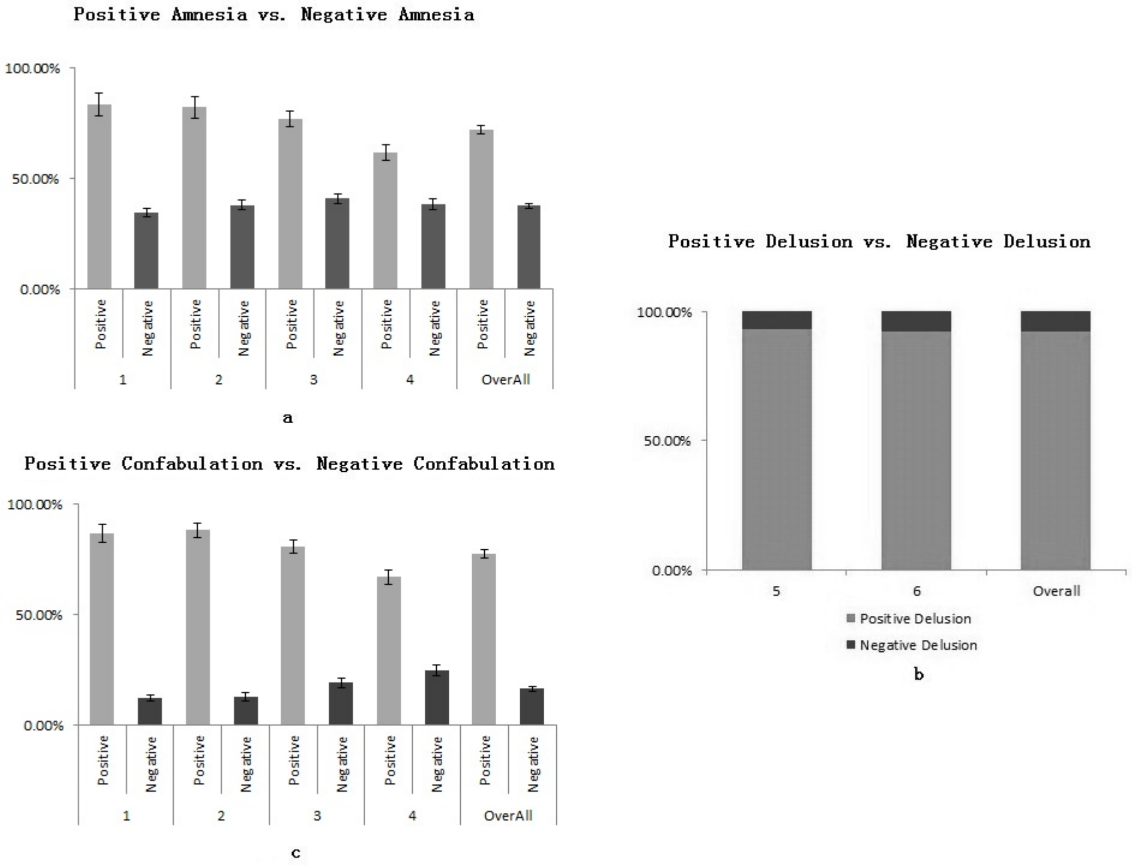


Figure 2: Memory Bias Conditional on Performance

respectively 93.3% and 91.7% in comparison with the base rates of correct response for questions 1 to 4: 85.0%, 81.1%, 64.2% and 56.3% (as shown in Table C2). Compared with Question 1, the rates of positive delusion for questions 5 and 6 are significantly higher than the base rate (respectively $p = 0.0000$ and $p = 0.0001$). All the other p values are at the 0.0000 level. Taken together, the pattern of delusion exhibits a significant positive tendency relative to the actual base rates of correct response. This behavior is compatible with the implications of our basic model but not for the Benabou-Tirole model.

Result 3 *Individuals tend to exhibit positive confabulation rather than negative confabulation.*

According to Table C1 in Appendix C, 48.9% of those who did it incorrectly in the first stage exhibit positive confabulation. As observed earlier. this is not compatible with our basic model. Figure 2c displays the respective rates of positive confabulation $a_B/(a_B + b_B)$ for Q1 to Q4 at 87.0%, 88.3%, 81.0% and 67.0%, which are significantly greater than the corresponding rates of negative confabulation $b_G/(a_G+b_G)$ at 11.9%, 12.6%, 19.0% and 24.5% (see Table C1). This behavior underscores the need for an extension of the basic model to incorporate positive confabulation in the next section.

We next observe:

Result 4 *Individuals who do not exhibit delusion are less likely to have positive confabulation; individuals with positive confabulation are more likely to have positive delusion.*

This result (see Table C4) describes the relationship between positive delusion and positive confabulation. Among 227 subjects out of 768 who do not exhibit positive delusion, the rates of positive confabulation of 25.0%, 15.2%, 19.8% and 21.3% for these four questions are significantly lower than the corresponding rates of unconditional positive confabulation of 52.2%, 57.2%, 49.6% and 43.6% at p-values of 0.0009, 0.0000, 0.0000 and 0.0000 respectively. Correspondingly, among 251 subjects who each exhibits some positive confabulation, their rates of positive delusion of 68.5.5% and 75.3% are significantly higher than the corresponding unconditional rates of positive delusion of 56.0% and 62.3% at p-values of 0.0002 and 0.0001 respectively. In sum, individuals who do not exhibit delusion are less likely to have positive confabulation while individuals with positive confabulation are more likely to have positive delusion. Together, these findings suggest that positive delusion may be an intermediate step in a process leading to positive confabulation. This 2-step confabulation process appears to be related to the idea of Korsakoff Syndrome (Whitty and Lewin, 1960) in which confabulation can be considered to serve as compensatory pseudo-reminiscence to fill the memory gap. In other words, the brain can produce false memory to make up for memory loss. We shall be building on this finding in the following section to develop an extension of our basic model.

3.2 Present Bias, IQ, and Memory Bias

When a subject does not recall whether he has seen a specific question previously, choosing (a), (b), or (c) entails some degree of downside risk or ambiguity. From this perspective, option (d) is free of risk or ambiguity. Before studying the implications of our model relating positive memory biases to the degree of present bias, we first examine whether the frequency of choosing (d) is related to risk attitude and ambiguity attitude measured using two incentivized tasks (see Appendix B). We run an ordered probit regression on the number of (d) choice, $\#d$ (from 1 to 6), with regressors β, IQ, RA, AA, and $Gender$, where IQ (out of 60) is the score on the Ravens test in the first stage, RA (from 0 to 10) refers to the observed degree of risk aversion in a portfolio choice task, AA (from -10 to 10) refers to the degree of ambiguity aversion, and *Gender* equals 1 if the subject is female and 0 otherwise. We do find a significant negative relation between $\#d$ with IQ which corroborates the general finding of a positive relation between IQ and accuracy of recall. There is no significant relation between $\#d$ and subjects' risk attitude, ambiguity attitude, or Gender. While the estimated coefficients for RA is negative, it is not significantly different from zero ($p = 0.102$).

In studying the influence of present bias, we first focus on positive memory biases based on subjects' recall of their performance on Q1 to Q4. Consider the subject's strategy for positive amnesia which applies when he did a question incorrectly. If he correctly recalls his performance in the second stage, i.e., choosing (b), the recall strategy is $h_B = 1$. If he cannot recall his initial performance or he does not remember having seen this question, i.e., choosing (c) or (d), we interpret this as positive amnesia with recall strategy $h_B = 0$. We run a probit regression on h_B with regressors β, IQ, RA, AA, and *Gender*.

Result 5 *Positive amnesia is not related to the degree of present bias.*

The estimated coefficients for β in questions 1 to 4 are of different signs and are individually not significant. Thus amnesia does not have a significant probit relation with present bias (see Appendix C.2 on amnesia). This is consistent with the implication of Proposition 2 but not with Proposition 1. Because amnesia in PBE3 is fake, the individual will recall the bad signal when β is either large enough or small enough and suppress the bad signal for intermediate values of β.

We next examine the subject's strategy for positive delusion with the two new questions – Q5 and Q6. If he indicates that he did it correctly, the delusion strategy is $h_\emptyset = 0$. Otherwise, if he indicates that this question may be new, i.e., answer (c) or (d), his delusion strategy is $h_\emptyset = 1$. We run a probit regression on $h_\emptyset$ using the regressors β, IQ, RA, AA, and *Gender*.

Result 6 *The likelihood of individual having positive delusion increases in the degree of present bias (β lower).*

For Question 5, the sign of the estimated coefficients for β is consistent with Corollary 1 and is significant ($p = 0.032$). The corresponding result for Question 6 is marginally significant ($p = 0.098$) with the same sign. Combining Q5 and Q6, the result of ordered probit regression is significant at $p = 0.036$ (see Appendix C.2 on delusion). This finding supports the implication of Corollary 1 of our basic model.

We now consider the subject's strategy for positive confabulation when he did a question incorrectly. If he indicates that he did it correctly, i.e., choosing (a), the confabulation strategy is $h'_B = 0$ which we interpret as positive confabulation. If he can recall his performance correctly, i.e., choosing (b), the confabulation strategy is $h'_B = 1$.

Result 7 *Positive confabulation is not related to the degree of present bias.*

We do not find a probit relation between confabulation and present bias. This finding is discussed further in the next section when we extend our basic model to account for the possibility of confabulation.

Finally, we examine the possible influence of IQ on memory patterns along with risk aversion, ambiguity aversion, and gender (see Appendix C.2). We first focus on unconditional memory patterns. Consistent with what has been reported in the literature, we find IQ to be positively related to accuracy in unconditional recall and actual performance for Q1 to Q4. Interestingly, in examining Q5 and Q6, besides being positively related to positive delusion, we find that IQ is positively related to unconditional delusion.

Next, we examine the relation between IQ memory bias in terms of positive amnesia, positive delusion and positive confabulation. Our findings are summarized below.

Result 8 *Higher IQ is positively related to a higher likelihood of positive delusion as well as positive confabulation, but not related to positive amnesia.*[10]

The sign of the estimated coefficients of positive delusion and positive confabulation for IQ are significantly negative for Q3 to Q6 (all at $p = 0.000$). For the two easiest questions (Q1 and Q2) with few getting them wrong, the estimated coefficient of confabulation is marginally significant ($p = 0.073$ and 0.031) with the same sign. This supports the implication of Proposition 10 that when p is lower (given that these two questions are relatively easy), a further lowering of the conditional probability of receiving a negative signal will increase the possibility of positive delusion, since the higher the IQ, the more stable the PBE3, leading in turn to more opportunities to be delusional. Our model is silent on the relation between IQ and positive confabulation.

[10] One may question if there may be an endogeneity problem since memory and IQ are known to be related. Here, our dependent variables in terms of amnesia, delusion, and confabulation are distinct from unconditional recall and forgetfulness for which our subjects' behavior accord with known findings in the literature. It seems implausible that better memory in terms of having a greater capacity for correct recall can lead to greater memory biases as summarized in Result 8. This observation also addresses a further question of reverse causality whether a greater capacity for correct memory may lead to higher IQ.

4 Extended Model with Confabulation

As discussed earlier, even though our basic model does not allow for confabulation, the incidence of confabulation accounts for a considerable proportion in the observed patterns of memory errors. Moreover, the observed confabulation also reveals a positive bias that the frequency of positive confabulation is significantly higher than the frequency of negative confabulation. We further observe in Result 4 that individuals exhibiting confabulation also tend to exhibit delusion. Taken together, our findings suggest a direction to extend our model by incorporating a 2-step process of positive confabulation – positive amnesia (bad to no signal) followed by positive delusion (no signal to good signal) – rather than for self-0 to directly transform a bad signal into a good signal (bad to good signal).

We posit a memory strategy involving four epochs: $t = 0, t = 0^+, t = 1$, and $t = 2$. At $t = 0$, the individual chooses his memory strategy after receiving a signal about his ability. At $t = 0^+$, the individual applies his memory strategy again after receiving the reported signal and further chooses how to transmit the reported signal – truthfully or deceptively. At $t = 1$, he decides whether to engage in the activity which delivers a payoff at $t = 2$. In this setting there are multiple ways to exhibit amnesia and delusion, given that motivated memory can happen at both $t = 0$ and $t = 0^+$.

Figure 3 displays the extended model with confabulation. Based on self-0^+'s memory, when $\hat{s} = B$, he can choose to recall it with probability $\widetilde{h_B}$ or forget it with probability $1 - \widetilde{h_B}$. When $\hat{s} = \emptyset$, he can choose to transmit it truthfully to self-2 with probability $\widetilde{h_\emptyset}$ or exhibit delusion with probability $1 - \widetilde{h_\emptyset}$. Accordingly, we define the reported signal transmitted from self-0^+ to self-1 as $\hat{\hat{s}}$. As before, self-1 will decide whether to undertake the task according to his belief about his ability.

Following the methodology in Section 2, we can identify 16 possible PBE's shown in Table 1 after applying the intuitive criterion (see Appendix A.11 for details). We classify the equilibria according to amnesia, delusion and confabulation by comparing the final signal $\widehat{\hat{s}}$ with the initial signal s. The first four equilibria are inherited from the basic model. In addition to PBE2, there are two additional amnesia equilibria consisting of PBE2a and PBE2b in the extended model. Similarly, besides PBE3a and PBE3b, there are six additional pure delusion equilibria: PBE3c to PBE3h. The extended model further delivers the sought after confabulation equilibria from PBE4a to PBE4d.

The key findings of how positive delusion and positive confabulation relate to the degree of present bias in the extended model are summarized below.

Proposition 12 *The likelihood of positive delusion and positive confabulation both increase in the degree of present bias.*

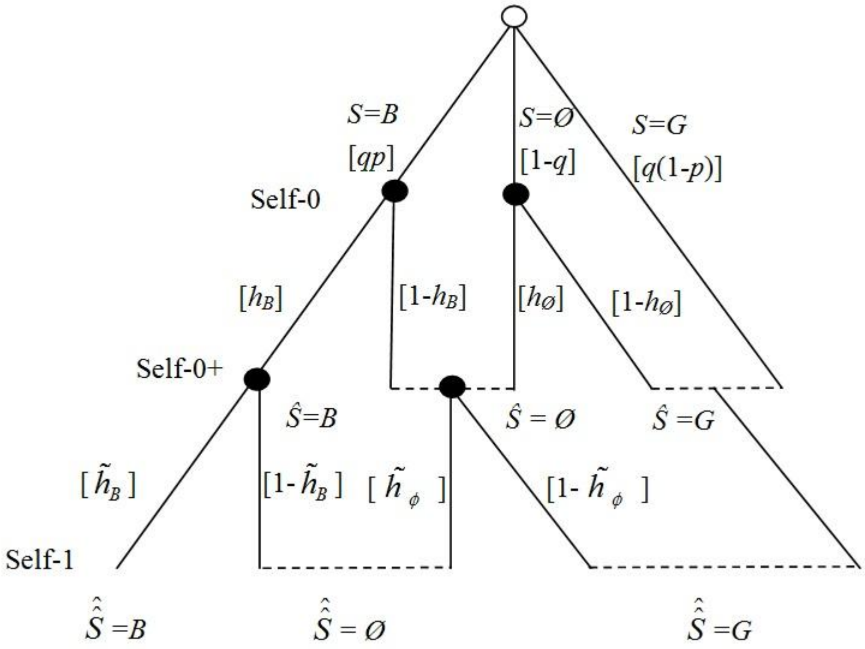
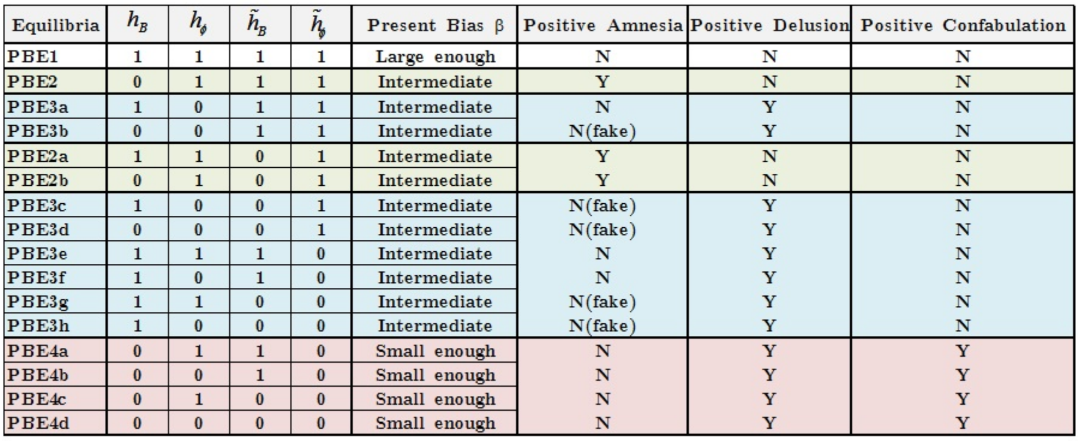


Figure 3: Two-Step Memory Strategy

Note that when self-0 receives a bad signal, self-1 will face three possibilities eventually: a bad signal, no signal and a good signal corresponding to perfect recall, amnesia and confabulation. When present bias is severe, individuals will exhibit positive memory bias regardless of whether he receives a bad signal or no signal in which case confabulation can appear without amnesia. This induces a motive for transforming a bad signal via confabulation into a good signal. In comparison with the basic model, our analysis here offers an alternative mechanism to deliver over confidence without involving amnesia.

While the additional implications of our extended model relating amnesia and delusion with present bias are supported by the experimental evidence presented in the preceding section, the corresponding relation between confabulation and present bias is not supported. A potential explanation may have to do with the duration between the two stages of the experiment. When β is sufficiently low, delusion always occurs, but the process of confabulation comprises amnesia as an additional step. Should the duration between the two stages be insufficient, this process may not progress beyond the first step, i.e., positive amnesia, which does not relate to present bias according to our data. Consequently, the incidence of confabulation may not be significant.



Equilibria	h_B	h_ϕ	$\tilde{h}_B$	$\tilde{h}_\phi$	Present Bias β	Positive Amnesia	Positive Delusion	Positive Confabulation
PBE1	1	1	1	1	Large enough	N	N	N
PBE2	0	1	1	1	Intermediate	Y	N	N
PBE3a	1	0	1	1	Intermediate	N	Y	N
PBE3b	0	0	1	1	Intermediate	N(fake)	Y	N
PBE2a	1	1	0	1	Intermediate	Y	N	N
PBE2b	0	1	0	1	Intermediate	Y	N	N
PBE3c	1	0	0	1	Intermediate	N(fake)	Y	N
PBE3d	0	0	0	1	Intermediate	N(fake)	Y	N
PBE3e	1	1	1	0	Intermediate	N	Y	N
PBE3f	1	0	1	0	Intermediate	N	Y	N
PBE3g	1	1	0	0	Intermediate	N(fake)	Y	N
PBE3h	1	0	0	0	Intermediate	N(fake)	Y	N
PBE4a	0	1	1	0	Small enough	N	Y	Y
PBE4b	0	0	1	0	Small enough	N	Y	Y
PBE4c	0	1	0	0	Small enough	N	Y	Y
PBE4d	0	0	0	0	Small enough	N	Y	Y

Table 1. Equilibrium Memory Bias for Extended Model

5 Discussion

In this section, we discuss further implications of our models in terms of consciousness defined on experience and mental well being from the decision making perspective, and offer a collectivist interpretation of our results.

5.1 Bounded Consciousness

As discussed in the Introduction, the assumption of full consciousness implicit in economic analysis involves unlimited attention, accurate encoding and storage of events, and perfect recall. Recently, Aumann (2006) suggests a definition of limited consciousness in terms of our bounded capacity to experience and offers his view that consciousness so defined is completely subjective, and yet it is the only phenomenon the observer may be certain of. Moreover, delusions, dreams, and ravings are themselves experiences; in each case, the observer is sure that what he is experiencing is conscious. Aumann also suggests an adaptive function for consciousness which begs the question of how memory bias may be adaptive in the social or natural environment. Ginsburg and Jablonka (2010) conjecture that adaptive memory may have emerged about 500 million years ago after the "Cambrian explosion" with the emergence of early nervous system giving the Cambrian fauna the capacity to adapt to their environment and memorize locations for food and water and signals for predators and mating partners. In humans, it has been shown that our neuronal circuits down to the synaptic levels are constantly being modified as we encode, store, and retrieve information (Bear, 2003) and that this mechanism of neuroplasticity operates throughout an individual's lifetime as discussed in Ansermet and Magistretti (2007). They offer evidence of neuroplasticity underpinning human unconsciousness which, in conjunction with Damasio's (1994) somatic marker hypothesis, may help direct decisions towards advantageous options nonconsciously. Our study of motivated memory biases demonstrates the possibility of using the tools of games theory and experimental economics to make inroads towards a better understanding of the role of bounded consciousness in decision making.

5.2 Mental Well Being

It has been argued that almost all modern humans today are descendents of the same small group of early humans who migrated out of Africa approximately only 85,000 years ago.[11] Following Diamond's (1997) demonstration that societies developed differently on different continents due to differences in natural environments, Ashraf and Galor (2012) offer an argument on how genetic diversity in relation to their proximity to Africa impacts in population density and productivity. This is in line with Todd and Gigerenzer's (2007) suggestion that simple heuristics for decision making may deliver better overall performance over an evolutionary time horizon in which human dispositions may be shaped by changes in social institutions and in biology. From this perspective, human biases or mental disorders may turn out to be functional (or endogenous) under more careful investigations (see, e.g., Foucault, 1961).

Over the past decades, worldwide suicide rates had increased significantly with the bulk of suicides being linked to mental disorder especially depression which may correspond to an extreme lack of motivation for life activities (Hawton and van Heeringen, 2009). In this regard, schizophrenia may reflect a tendency of delusion to compensate for a lack of motivation. Sass (1998) reports the interplay of two opposing experiences among schizophrenics. Many schizophrenic patients lose their sense of active intentionality which we may interpret as self doubt or under confidence (self-1 with $s = \widehat{s} = G$). At the same time, they may experience over confidence or euphoria of being preeminent (self-1 with $s = \emptyset$, $\widehat{s} = G$). This behavior may correspond to PBE3 with delusion in Proposition 2. More importantly, our key theoretical finding in Corollary 1 and experimental finding in Result 6 are compatible with the report of comorbidity between present bias and the onset of schizophrenia (Heereya, Robinsona, McMahona and Golda, 2007). On top of this, Proposition 8 suggests that the incidence of mental disorder involving amnesia and schizophrenia would increase with a greater degree of present bias when the signal is more informative (high γ in our model). Furthermore, the implication of Proposition 9 accords with recent evidence

[11] See, e.g., Roberts (2009) for a detailed review, Atkinson (2011) for the related recent evidence in linguistics, and Ashraf and Galor (2011) for its relevance for economic development.

on a positive relation between onset of schizophrenia and social isolation (low q) in Jones, et al (1994) which points out that children who tend to exhibit social isolation are more likely to suffer from schizophrenia as adults. There are also corroborating evidence about individuals with autism having a greater risk for schizophrenia (see, e.g., Russell, Bott and Sammons, 1989; Watkins, Asarnow and Tanguay, 1988) whose incidence is associated with a lack of social interactions. Building on the functionality and adaptive consequences of motivated memory and delusion, our method and results may contribute to recent attempts to understand mental well being from the decision making perspective using multiple platforms including neuroimaging, genetics, and pharmacological intervention (see review and discussion in Sawa and Snyder, 2002; Chiu et. al., 2008; Kishida, King-Casas, and Montague, 2010).

5.3 Collectivist Interpretation

As has been attempted in the literature, our model admits natural reinterpretation in a multi-person setting requiring no substantive change in game form. Such a reinterpretation may provide a rationale for the motivational value of myth making, e.g., telling fairy tales to induce children to form a more rosy view of the world corresponding to a belief in an enhanced chance of future happiness. Proposition 2 suggests that telling tales may conceivably be more functional than omitting information such as a lack of academic achievements. In this case, self-0 represents the older generation while self-1 refers to the younger generation. Furthermore, besides present bias at the individual level, the β parameter may also capture a degree of altruism of the current generation towards the future generation.[12] Here, forgetfulness has the natural interpretation as the collective amnesia of the younger generation. For example, older generations in Japan could revise historical texts by down playing the event of 'Nanking massacre' for the young.[13] Moreover, our extended model can be applied to model collective confabulation in transforming past disastrous events into myths, legends, and Utopian tales to be transmitted across generations, e.g., the official account of China's great leap forward in which the younger generation is taught about its many virtues. Local leaders falsely reported high grain production figures as they witnessed mass starvation and famine (see, e.g., Yang et. al., 2012). Our approach can give rise to a fresh take on how institutional fabrication including collective am-

[12] In his seminal work, Simon (1976) posits that young people have a capacity which he terms "docility" on which their parents and elders may instill tastes, habits and values thereby shaping their identity. This gives rise to a mechanism for parents and elders to take a substantive interest in the well being of their young giving rise to a older-generation bias that is akin to present bias.

[13] In this respect, closest to our theory is the work by Dessi (2008) who studies collective memory and cultural transmission, and explains how information suppression at the societal level alleviates the free riding problem.

nesia, collective delusion, and collective confabulation can enhance confidence at the national level, thereby motivating people to invest in the collective good from the perspective of the older generations.

6 Concluding Remarks

The full rationality assumption implicit in the fabled homo economicus has served as the work horse for much of economic analysis. Traditionally, the homo economicus has been endowed with a narrow sense of self interest both in terms of exhibiting little regard for the well being of others and having an exclusive focus on consequences regardless of how they may arise. Both these aspects of full rationality have been challenged by evidence from a huge and growing literature in behavioral and experimental economics, e.g., people are prepared to engage in costly activities for the benefit of others or out of sense of fairness. Beyond distributional considerations, how consequences arise may also matter (Rabin, 1998; Falk, Fehr, and Fischbacher, 2003; Camerer, Loewenstein and Rabin, 2003). Another dimension of full rationality concerns an unlimited ability to compute, to reason logically, and to infer accurately and update with ease and without bias. Since the work of Simon (1947), idea that actual decision makers have limited computational ability there has been increasingly recognized (see Kahneman, Slovic, and Tversky, 1982; Rubinstein, 1998).

Our paper addresses the need to moderate a third and less studied dimension of the homo economicus, namely full consciousness. We develop a model of limited consciousness by focusing on the role of selective memory and motivated delusion and test their implications in a controlled and incentivized experimental setting. Our formulation enables us to probe more deeply into the cognitive processes governing the decision maker's possibly nonconscious memory biases. By reinterpreting different epochs as stages in an atemporal multi-stage decision making setting, our approach also lends itself to being adapted to model selective (in)attention allowing for the possibility of fantasies.

A natural follow up question concerns whether motivated memory may relate to additional temporal decision making traits beyond discounting, e.g., anxiety in terms of the timing of uncertainty resolution (Kreps and Porteus, 1978; Chew and Epstein, 1989; Chew and Ho, 1994; LoVallo and Kahneman, 2000; Grant, Kajii, and Polak, 1998) and risky decision making traits such as Allais behavior, longshot preference, and familiarity bias. In a similar vein, it would also be natural to study the potential influence of cognitive and personality traits as well as judgment related biases (see, e.g., Tversky and Kahneman, 1977; Rabin and Schrag, 1999).

Another follow up question concerns how motivated memory may relate to moral sentiments and other regarding behavior. In this regard, the individual's social cognition skills will presumably be called into question, e.g., how ordinary people feel and think about their society (see, e.g., Fiske and Taylor, 2007). Thus, a wide range of cognitive processes deserve our attention as we attempt to identify the condition under which specific ones will prevail. At the collective level, it is well known in cultural psychology and social anthropology that self-cognitions are different across cultures. In the economics literature, for example, Dessi and Zhao (2011) endogenize cultural differences on self-esteem, social emotions, and self discipline. At the individual level, the meaning of mental (dis)order may be distinct for people living under different environments.

This paper is part of an evolving literature modeling aspects of human cognition and consciousness as they relate to economic decision making. To varying extent, we have discussed biological factors alongside psychological considerations. The stage appears to be set for bringing biology including neuroscience and molecular genetics more fully into the modelling of motivated cognitive processes in studying bounded consciousness in decision making.

References

[1] Arai, J.A., S. Li, M.D. Hartley, and L.A. Feig (2009). "Transgenerational Rescue of a Genetic Defect in Long-Term Potentiation and Memory Formation by Juvenile Enrichment," *Journal of Neuroscience*, 29(5), 1496 –1502.

[2] Ashraf, Q. and O. Galor (2011). "The Out of Africa Hypothesis, Human Genetic Diversity and Comparative Development," *American Economic Review*, forthcoming.

[3] Atkinson, Q. D. (2011). "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa," *Science*, 332, 346-349.

[4] Aumann, R. J. (2006). "Consciousness," *Life as We Know It*, edited by J. Seckbach, Springer, Dordrecht, 555-564.

[5] Batel, P. (2000). "Addiction and Schizophrenia," *European Psychiatry*, 15, 115-122.

[6] Benabou, R. and J. Tirole (2002). "Self-Confidence and Personal Motivation," *Quarterly Journal of Economics,* 117(3), 871-915.

[7] Benoit, J. P. and J. Dubra (2011). "Apparent Overconfidence," *Econometrica*, 79(5),1591-1625.

[8] Beste, C., C. Saft, O. Guentuerkuen and M. Falkenstein (2008). “Increased Cognitive Functioning in Symptomatic Huntington’s Disease as Revealed by Behavioral and Event-Related Potential Indices of Auditory Sensory Memory and Attention,” *Journal of Neuroscience*, 28(45), 11695-11702.

[9] Blake, D.T., N.N. Byl, M.M. Merzenich (2002). “Representation of the Hand in the Cerebral Cortex,” *Behavioural Brain Research*, 135(1-2), 179-84.

[10] Bortolotti, M. and M. Mameli (2012). “Self-Deception, Delusion and the Boundaries of Folk Psychology,” *Humanamente,* 20, 203–221.

[11] Brown, T. L., R. T. A. Croson, and C. Eckel (2011). “Intra- and Inter-personal Strategic Ignorance: A Test of Carrillo and Mariotti,” working paper.

[12] Burks, S. V., J. P. Carpenter, L. Goette and A. Rustichini (2012). “Overconfidence and Social-Signaling,” *Review of Economic Studies,* forthcoming.

[13] Butler, P. (2000). “Reverse Othello Syndrome Subsequent to Traumatic Brain Injury,” *Psychiatry*, 63, 85-92.

[14] Camerer, C., G. Loewenstein, and M. Rabin (2003). *Advances in Behavioral Economics*, Princeton University Press.

[15] Carrillo, J. and T. Mariotti (2000). “Strategic Ignorance as a Self-Disciplining Device,” *Review of Economic Studies*, 67(3), 529-44.

[16] Chew, S. H. and L. Epstein, (1989). “The Structure of Preferences and Attitudes Towards the Timing of Resolution of Uncertainty,” *International Economic Review*, 30, 103–117.

[17] Chew, S. H. and J. L. Ho, (1994). “Hope: An empirical study of attitude toward the timing of uncertainty resolution,” *Journal of Risk and Uncertainty*, 8, 267–288.

[18] Chiu, P.H., Kayali, M.A., Kishida, K.T., Tomlin, D., Klinger, M.R., and Montague, P.R. (2008). “Self Responses along Cingulate Cortex Reveal Quantitative Neural Phenotype for High-Functioning Autism,” *Neuron*, 57, 463–473.

[19] Cho, I. K. and D. M. Kreps, (1987). “Signaling Games and Stable Equilibria,” *Quarterly Journal of Economics*, 102, 179-221.

[20] Cochran, G., J. Hardy and H. Harpending (2005). “Natural History of Ashkenazi Intelligence,” *Journal of Biosocial Science*, 38, 1-35.

[21] de Cervantes, M. (1605). *Don Quixote*, (translated by C. Jarvis), Oxford University Press, 1992.

[22] Dessi, R. (2008). "Collective Memory, Cultural Transmission and Investments," *American Economic Review* 98(1), 534-560.

[23] Dessi, R. and X. Zhao (2011). "Self-Esteem, Shame and Personal Motivation," IDEI Working Paper, n. 639.

[24] Diamond, J. (1997). *Guns, Germs and Steel: The Fates of Human Societies*, W.W. Norton.

[25] Dow, J. (1991). "Search Decisions with Limited Memory," *Review of Economic Studies*, 58(1), 1-14.

[26] Dunning, D. (2001). "On the Motives Underlying Social Cognition," in A. Tesser and N. Schwarz (eds.) *Blackwell Handbook of Social Psychology: Intraindividual Processes.*

[27] Falk, A., Fehr, E. and U. Fischbacher (2003). "On the Nature of Fair Behavior," *Economic Inquiry*, 41, 20-26.

[28] Fiske, S. and S. Taylor (2007). *Social Cognition, from Brains to Culture*, McGraw-Hill, 2007.

[29] Foucault, M. (1961). *Madness and Civilization: A History of Insanity in the Age of Reason*, (translated by R. Howard) Vintage, New York, 1988.

[30] Furman, L. (2005). "What Is Attention-Deficit Hyperactivity Disorder (ADHD)?" *Journal of Child Neurology*, 20, 994-1002.

[31] Ginsburg, S. and E. Jablonka (2010). "The Evolution of Associative Learning: A factor in the Cambrian explosion," *Journal of Theoretical Biology*, 266(1), 11-20.

[32] Gottlieb, D. (2010). "Imperfect Memory And Choice Under Risk," mimeo, University of Pennsylvania.

[33] Grant, S., A. Kajii., and B. Polak, (1998). "Intrinsic Preference for Information," *Journal of Economic Theory*, 83, 233-259.

[34] Guillem, F., T. Pampoulova, E. Stip, C. Todorov and P. Lalonde (2005). "Are There Common Mechanisms In Sensation Seeking And Reality Distortion In Schizophrenia? A Study Using Memory Event-Related Potentials," *Psychiatry Research*, 135(1), 11-33.

[35] Hawton, K. and K. van Heeringen (2009). “Suicide,” *Lancet*, 373, 1372–81.

[36] Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory* New York: Wiley, 1949.

[37] Heereya, E.A., B.M. Robinsona, R.P. McMahona and J.M. Gold (2007).“Delay Discounting in Schizophrenia,” *Cognitive Neuropsychiatry*, 12(3), 213-221.

[38] Hemsley, D., J. Rawlins, J. Feldon, S. Jones, J. Gray (1993). “The Neuropsychology of Schizophrenia: Act 3,” *Behavioral and Brain Sciences*, 16, 209-215.

[39] Howe, M. L. (2011). “The Adaptive Nature of Memory and Its Illusions,” *Psychological Science*, 20(5), 312-315.

[40] Howe, M. L. and M. H. Derbish (2010). “On The Susceptibility of Adaptive Memory To False Memory Illusions,” *Cognition*, 115(2), 252–267.

[41] Howe, M.L., S. R. Garner, M. Charlesworth, and L. Knott, (2011). “A Brighter Side To False Memory Illusions: False Memories Can Prime Children’s And Adults’ Insight-based Problem Solving,” *Journal of Experimental Child Psychology*, 108, 383–393.

[42] Jones, P., R. Murray, P. Jones, B. Rodgers and M. Marmot (1994).“Child Developmental Risk Factors For Adult Schizophrenia In The British 1946 Birth Cohort,” *Lancet*, 334, 1398-1402.

[43] Kahneman, D., P. Slovic, and A. Tversky, (1982). *Judgment Under Uncertainty: Heuristics and biases.* New York: Cambridge University Press.

[44] Kandel, E.R. (2001). “The Molecular Biology of Memory Storage: A Dialogue Between Genes and Synapses,” *Science,* 294 (5544), 1030-1038.

[45] Kishida, K. T., B. King-Casas, and P. R. Montague. (2010). “Neuroeconomic Approaches to Mental Disorders,” *Neuron*, 67, 543–554.

[46] Kreps, D. M., and E. L. Porteus, (1978). “Temporal Resolution of Uncertainty and Dynamic Choice Theory,” *Econometrica*, 46(1), 185-200.

[47] Knudsen, E. (2007). “Fundamental Components of Attention,” *Annual Review of Neuroscience*, 30, 57-78.

[48] Koszegi, B. (2006). “Ego Utility, Overconfidence, and Task Choice,” *Journal of the European Economic Association*, 4(4), 673-707.

[49] Laibson, D. (1997). "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics,* 62, 443-477.

[50] Li, K. (2012). "Asymmetric Memory Recall of Positive and Negative Events in Social Interactions," *Experimental Economics*, forthcoming.

[51] Lovallo, D. and D. Kahneman , (2000). "Living with Uncertainty: Attractiveness and Resolution Timing," *Journal of Behavioral Decision Making*, 13, 179-190.

[52] Lyon, V. and M. Fitzgerald (2005). *Asperger Syndrome: a Gift or a Curse?* Nova Science, 2005.

[53] Malenka, R. C. (2003). "Synaptic Plasticity and AMPA Receptor Trafficking," *NYPAS*, 1003, 1-11.

[54] McKay, R., R. Langdon and M. Coltheart, (2005). "Paranoia, Persecutory Delusions and Attributional Biases," *Psychiatry Research*, 136(2-3), 233-245.

[55] McKay, R. T. and D. C. Dennett (2009). "The Evolution of Misbelief," *Behavioral and Brain Sciences*, 32, 493-561.

[56] Meshi D., M.R. Drew, M. Saxe, M.S. Ansorge, D. David, L. Santarelli, C. Malapani, H. Moore, and R. Hen (2006). "Hippocampal Neurogenesis is Not Required for Behavioral Effects of Environmental Enrichment," *Nature Neuroscience*, 9, 729-731.

[57] Pashler, H. (1998). *The Psychology of Attention*, MIT Press, Cambridge, MA.

[58] Rabin, M. (1998). "Psychology and Economics," *Journal of Economic Literature,* 36, 11–46.

[59] Rabin, M and J L Schrag (1999). "First Impressions Matter: A Model of Confirmatory Bias," *Quarterly Journal of Economics,* 114(1), 37-82.

[60] Ramachandran, V. S. (1996). "The Evolutionary Biology of Self-deception, Laughter, Dreaming and Depression: Some Clues From Anosognosia," *Medical Hypotheses*, 47(5), 347-362.

[61] Roberts, A. (2009). *The Incredible Human Journey*, Bloomsbury, 2009.

[62] Rubinstein, A. (1998). *Modeling Bounded Rationality*, MIT Press.

[63] Russell, A., L. Bott, and C. Sammons (1989). "The Phenomenology of Schizophrenia Occurring in Childhood," *Journal of the American Academy of Child & Adolescent Psychiatry*, 28(3), 399-407.

[64] Sass, A. (1998). “Schizophrenia, Self-Consciousness, and the Modern Mind,” *Journal of Consciousness Studies*, 5(5-6), 543-565.

[65] Sawa, A. and S. Snyder (2002). “Schizophrenia: Diverse Approaches to a Complex Disease,” *Science*, 296, 692-695.

[66] Simon, H. (1955). “A Behavioral Model of Rational Choice,” *Quarterly Journal of Economics*, 69, 99–118.

[67] Simon, H. (1947). *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization,* Macmillan and subsequent edition in 1976.

[68] Sims, C. A. (2003). “Implications of Rational Inattention,” *Journal of Monetary Economics*, 50(3), 665–690.

[69] Strotz, R. (1955). “Myopia and Inconsistency in Dynamic Utility Maximization,” *Review of Economic Studies,* 23, 165-180.

[70] Svenson, O. (1981). “Are We All Less Risky and More Skillful Than Our Fellow Drivers?” *Acta Psychologica*, 94, 143-148.

[71] Todd, M. and G. Gigerenzer (2007). “Environments that Make Us Smart: Ecological Rationality,” *Current Directions in Psychological Science*, 16, 167-171.

[72] Tversky, A. and D. Kahneman (1977). “Casual Schema in Judgments Under Uncertainty,” in Fishbein, Martin (Ed.), Progress in Social Psychology, Hillsdale: Erlbaum.

[73] Watkins, J., R. Asarnow, and P. Tanguay (1988). “Symptom Development In Childhood Onset Schizophrenia,” *Journal of Child Psychology and Psychiatry*, 29(6), 865–878.

[74] Weinberg, B. A. (2006). “A Model of Overconfidence,” mimeo, Ohio State University.

[75] Whitty, C. W. M. and W. A. Lewin. (1960). “Korsakoff Syndrome in the Postcingulectomy Confusional State,” *Brain*, 83, 648-653.

[76] Williams, B.M., Y. Luo, C. Ward, K. Redd, R. Gibson, S.A. Kuczaj, J.G. McCoy (2001). “Environmental Enrichment: Effects on Spatial Memory and Hippocampal CREB Immunoreactivity,” *Physiology and Behavior*, 73, 649–658.

[77] Yang, J., E. Friedman, S. Mosher, G. Jian, R. MacFarquhar (2012). *Tombstone: The Great Chinese Famine, 1958-1962,* Farrar, Straus and Giroux.

Appendix

A Proof of Propositions

A.1 Proof of Proposition 2

(i) *PBE1* ($h_B^* = 1$, $h_\emptyset^* = 1$): We have $r^*(\emptyset) = 1$ and $r^*(G) = 1$. When $s = B$, define $\chi[r^*(\emptyset), \beta] = U_S(\theta_B) - U_T(\theta_B)$. This is given by:

$$\int_{\beta\theta_B V}^{\beta\theta^*(\emptyset)V} \{\theta_B V - c\} dF(c) = \int_{\beta\theta_B V}^{\beta\theta_\emptyset V} \{\theta_B V - c\} dF(c).$$

Notice that $\chi[r^*(\emptyset), 1] < 0$ and that $\chi[r^*(\emptyset), \beta] > 0$ for $\beta \in (0, \theta_B/\theta^*(\emptyset)]$. It follows that there exists $\beta' \in (\theta_B/\theta^*(\emptyset), 1)$ such that $\chi[r^*(\emptyset), \beta'] = 0$. Moreover, $\chi[r^*(\emptyset), \beta]$ is positive for $\beta \in (0, \beta')$ and negative for $\beta \in (\beta', 1)$, and

$$\frac{\partial \chi[r^*(\emptyset), \beta]}{\partial \beta} = \theta^*(\emptyset)V^2[\theta_B - \beta\theta^*(\emptyset)]f[\beta\theta^*(\emptyset)V] - \theta_B V^2[\theta_B - \beta\theta_B]f(\beta\theta_B V) < 0,$$

for $\beta \in [\theta_B/\theta^*(\emptyset), 1]$. Therefore, $h_B^* = 1$, if $\beta > \beta'$.

When $s = \emptyset$, we can similarly define $\Psi[r^*(\emptyset), r^*(G), \beta] \equiv U_F(\theta_\emptyset) - U_T(\theta_\emptyset)$ which is given by:

$$\int_{\beta\theta^*(\emptyset)V}^{\beta\theta^*(G)V} \{\theta_\emptyset V - c\} dF(c) = \int_{\beta\theta_\emptyset V}^{\beta\theta_G V} \{\theta_\emptyset V - c\} dF(c).$$

We have that $\Psi[r^*(\emptyset), r^*(G), 1] < 0$ and $\Psi[r^*(\emptyset), r^*(G), \beta] > 0$, for $\beta \in (0, \theta_\emptyset/\theta_G]$. Thus, $\Psi[r^*(\emptyset), r^*(G), \beta''] = 0$ for some $\beta'' \in (\theta_\emptyset/\theta_G, 1)$. Moreover, $\Psi[r^*(\emptyset), r^*(G), \beta]$ is positive for $\beta \in (0, \beta'')$ and negative for $\beta \in (\beta'', 1)$, and

$$\frac{\partial \Psi[r^*(\emptyset), r^*(G), \beta]}{\partial \beta} = \theta_G V^2[\theta_\emptyset - \beta\theta_G]f(\beta\theta_G V) - \theta_\emptyset V^2[\theta_{B\emptyset} - \beta\theta_\emptyset]f(\beta\theta_\emptyset V) < 0,$$

for $\beta \in [\theta_\emptyset/\theta_G, 1]$. Thus, $h_\emptyset^* = 1$, if $\beta > \beta''$.

It follows that a correct recall PBE1 ($h_B^* = 1$, $h_\emptyset^* = 1$) exists for $\beta > \max\{\beta', \beta''\}$.

(ii) *PBE2* ($h_B^* = 0$, $h_\emptyset^* = 1$): We have $r^*(\emptyset) = (1-q)/(qp + 1 - q)$ and $r^*(G) = 1$. When $s = B$,

$$\chi[r^*(\emptyset), \beta] = \int_{\beta\theta_B V}^{\beta[\frac{1-q}{qp+1-q}\theta_\emptyset + \frac{qp}{qp+1-q}\theta_B]V} \{\theta_B V - c\} dF(c).$$

As with the proof of existence of PBE1, there exists $\overline{\beta}_2$ such that $\chi[r^*(\emptyset), \beta] > 0$ for $\beta \in (0, \overline{\beta}_2)$ and $\chi[r^*(\emptyset), \beta] < 0$ for $\beta \in (\overline{\beta}_2, 1)$.It follows that $h_B^* = 0$ if $\beta < \overline{\beta}_2$.

When $s = \emptyset$,

$$\Psi[r^*(\emptyset), r^*(G), \beta] = \int_{\beta(\frac{1-q}{qp+1-q}\theta_\emptyset + \frac{qp}{qp+1-q}\theta_B)V}^{\beta\theta_G V} \{\theta_\emptyset V - c\} dF(c).$$

We can similarly conclude that there is a $\underline{\beta}_2$ such that $\Psi[r^*(\emptyset), r^*(G), \underline{\beta}_2] = 0$ so that $h^*_\emptyset = 1$ for $\beta > \underline{\beta}_2$ since $\Psi[r^*(\emptyset), r^*(G), \beta] > 0$ for $\beta \in (0, \underline{\beta}_2)$ and $\Psi[r^*(\emptyset), r^*(G), \beta] < 0$ for $\beta \in (\underline{\beta}_2, \theta_\emptyset/\theta^*(\emptyset))$.

The existence of PBE2 requires $0 < \underline{\beta}_2 < \overline{\beta}_2 < 1$. Under the assumption of uniform distribution, these two thresholds can be solved explicitly:

$$\overline{\beta}_2 = \frac{2(1-q+qp)\theta_B}{(1-q+p+qp)\theta_B + (1-p)(1-q)\theta_G}$$

and

$$\underline{\beta}_2 = \frac{2[1+(p-1)q][p(\theta_B - \theta_G) + \theta_G]}{2(1-q)\theta_G + p[\theta_B + (2q-1)\theta_G]}.$$

It follows that $1/2 < p < 1$ and $1 < \gamma < p(p-q+qp)/(1-p)^2(1-q)$, which can be satisfied for sufficiently large p. Thus PBE2 ($h^*_B = 0$, $h^*_\emptyset = 1$) exists if $\underline{\beta}_2 < \beta < \overline{\beta}_2$.

(iii) *PBE3a* ($h^*_B = 1$, $h^*_\emptyset = 0$): We have $r^*(G) = (q-qp)/(1-qp)$ with $r^*(\emptyset)$ arbitrary since it is an off the equilibrium path belief. When $s = B$,

$$\chi[r^*(\emptyset), \beta] = \int_{\beta\theta_B V}^{\beta[r^*(\emptyset)\theta_\emptyset + (1-r^*(\emptyset))\theta_B]V} \{\theta_B V - c\} dF(c).$$

Notice that $\chi[r^*(\emptyset), \beta] = 0$ and $h^*_B = 1$ when $r^*(\emptyset) = 0$. If $r^*(\emptyset) > 0$, we can show similarly that there exists a $\underline{\beta}_3$ such that $\chi[r^*(\emptyset), \underline{\beta}_3] = 0$ with $\chi[r^*(\emptyset), \beta]$ positive or negative depending on whether β is less than or greater than $\underline{\beta}_3$. Thus $h^*_B = 1$, if $\beta > \underline{\beta}_3$.

Similarly, when $s = \emptyset$,

$$\Psi[r^*(\emptyset), r^*(G), \beta] = \int_{\beta[r^*(\emptyset)\theta_\emptyset + (1-r^*(\emptyset))\theta_B]V}^{\beta[\frac{q-qp}{1-qp}\theta_G + \frac{1-q}{1-qp}\theta_\emptyset]V} \{\theta_\emptyset V - c\} dF(c).$$

We can show that there is a $\overline{\beta}_3$ solving $\Psi[r^*(\emptyset), r^*(G), \overline{\beta}_3] = 0$ such that $\Psi[r^*(\emptyset), r^*(G), \beta]$ is positive or negative depending on whether β is less than $\overline{\beta}_3$ or greater than $\overline{\beta}_3$ but less than $\theta_\emptyset/\theta^*(\emptyset)$. Thus, $h^*_\emptyset = 0$ if $\beta < \overline{\beta}_3$. It follows that PBE3a ($h^*_B = 1$, $h^*_\emptyset = 0$) exists if $\underline{\beta}_3 < \beta < \overline{\beta}_3$.

Now we consider the intuitive criteria to refine this equilibrium. Note that this equilibrium is determined by the off-equilibrium belief $r^*(\emptyset)$. When $r^*(\emptyset) = 0$, self-B is indifferent between recall and amnesia. When $r^*(\emptyset) > 0$, he will choose to recall the bad signal for sufficiently large β. For type-$\emptyset$ self, regardless of the value of $r^*(\emptyset)$, delusion is always strictly better than correct recall when $\beta < \overline{\beta}_3$. Thus after the equilibrium refinement under the intuitive criterion, self-1 knows that self-0 of type-$\emptyset$ will not correctly recall not having received a signal. Receiving such an empty signal precludes being type-$\emptyset$, so that the off-equilibrium-path belief $r^*(\emptyset)$ can only be zero. In this case, the decision maker is indifferent between correct recall and amnesia and would not need a large β to remain self truthful. Thus, when β is small enough, this equilibrium prevails.

PBE3b (*$h^*_B = 0$, $h^*_\emptyset = 0$*): We have $r^*(\emptyset) = 0$ and $r^*(G) = (q - qp)/(1 - qp)$. When $s = B$,

$$\chi[r^*(\emptyset), \beta] = \int_{\beta\theta_B V}^{\beta\theta_B V} \{\theta_B V - c\} dF(c) = 0.$$

Thus $\chi[r^*(\emptyset), \beta] = 0$, i.e., self-$B$ has no incentive to deviate from suppressing the bad signal for any β.

When $s = \emptyset$,

$$\Psi[r^*(\emptyset), r^*(G), \beta] = \int_{\beta\theta_B V}^{\beta[\frac{q-qp}{1-qp}\theta_G + \frac{1-q}{1-qp}\theta_\emptyset]V} \{\theta_\emptyset V - c\} dF(c).$$

Similarly, we can show that there exists $\overline{\beta}_3$ such that $\Psi[r^*(\emptyset), r^*(G), \overline{\beta}_3] = 0$ and that $\Psi[r^*(\emptyset), r^*(G), \beta]$ is positive for $\beta \in (0, \beta_3)$ and is negative for $\beta \in (\beta_3, \theta_\emptyset/\theta_B]$. Therefore $h^*_\emptyset = 0$ if $\beta < \beta_3$. it follows that PBE3b ($h^*_B = 0$, $h^*_\emptyset = 0$) exists if $\beta < \beta_3$.

Q.E.D.

A.2 Proof of Proposition 3

The existence of PBE2 requires that $\underline{\beta}_2 < \overline{\beta}_2$ which derives that $1/2 < p < 1$ and $1 < \gamma < p(p + (p-1)q) / \left[(1-p)^2 (1-q)\right]$.

We firstly check the relation between β_1 and $\overline{\beta}_2$. We have:

$$\begin{aligned}\overline{\beta}_2 - \beta' &= \frac{2(1 - q + qp)}{(1 - q + p + qp) + (1 - p)(1 - q)\gamma} - \frac{2}{(1 + p) + (1 - p)\gamma} \\ &= \frac{2(\gamma - 1)(1 - p)qp}{[1 + \gamma(1 - p) + p][1 - q + p + \gamma(1 - p)(1 - q) + qp]} > 0\end{aligned}$$

and

$$\begin{aligned}\overline{\beta}_2 - \beta'' &= \frac{2(1 - q + qp)}{(1 - q + p + qp) + (1 - p)(1 - q)\gamma} - \frac{2[p(1 - \gamma) + \gamma]}{p(1 - \gamma) + 2\gamma} \\ &= \frac{2(\gamma - 1)[p^2 - \gamma(1 - p)^2(1 - q)]}{[\gamma(2 - p) + p][1 + p + \gamma(1 - p)(1 - q) - q + qp]}\end{aligned}$$

of which the sign is determined by the term $p^2 - \gamma(1-p)^2(1-q)$ that is decreasing in γ. When γ reaches its minimum, the term is equal to $(1-p)pq$ which is always positive. Thus we can conclude that $\overline{\beta}_2 - \beta_1 > 0$.

Then we check the relation between $\underline{\beta}_2$ and β_3.We will study the monotonicity of the thresholds systematically in the next part. Here we just borrow the result that $\underline{\beta}_2$ is increasing in q while β_3 is decreasing in q. When $q = 1$, we can get:

$$\beta_3 = \underline{\beta}_2 = \frac{2[\gamma - (\gamma-1)p]}{1+\gamma}$$

Thus when $0 < q < 1$, we have $\beta_3 > \underline{\beta}_2$.

The overlapping between PBE1 and PBE2 requires that $\overline{\beta}_2 - \beta' > 0$ and $\overline{\beta}_2 - \beta'' > 0$ which always hold. The first inequality can be insured by the complementary belief of action in our model. The action of self-0 of type-B h_B is positive related to self-1 of type-B's belief: the more self-1 trust self-0 ($r^*(\emptyset)$ is higher), the more likely self-0 will choose to tell the truth (h_B is more likely to be 1), and vice versa, so the double equilibria emerge. The second inequality requires the existence of PBE2. The difference between two critical values can be insured to be positive for a minimal γ. The difference is decreasing in γ and is thus always positive. The overlapping between PBE2 and PBE3 is straightforward: the two critical values are the same when q reaches its maximum. The opposite monotonicity of the two critical values insures the overlapping of them.

Q.E.D.

A.3 Proof of Proposition 4

If $\underline{\beta}_2 > \overline{\beta}_2$, PBE2 does not exist. This derives the following conditions: either $0 < p \leqslant 1/2$, or $1/2 < p < 1$ and $\gamma > p(p + (p-1)q) / \left[(1-p)^2(1-q)\right]$. Now we check the relation between β_1 and β_3. Borrowing the results of the next part, we have β_3 is decreasing in q, while β_1 is independent of q. Now assume $q = 1$, we have:

$$\begin{aligned}\beta_3 - \beta' &= \frac{2(pq-1)[p(1-\gamma)+\gamma]}{[p(2q-1)-1]+(p-1)\gamma} - \frac{2}{(1+p)+(1-p)\gamma} \\ &= \frac{2(\gamma-1)(1-p)[1-p-p^2+\gamma(1-p)^2]}{(1+\gamma)(1-p)[(1+p)+(1-p)\gamma]}\end{aligned}$$

and

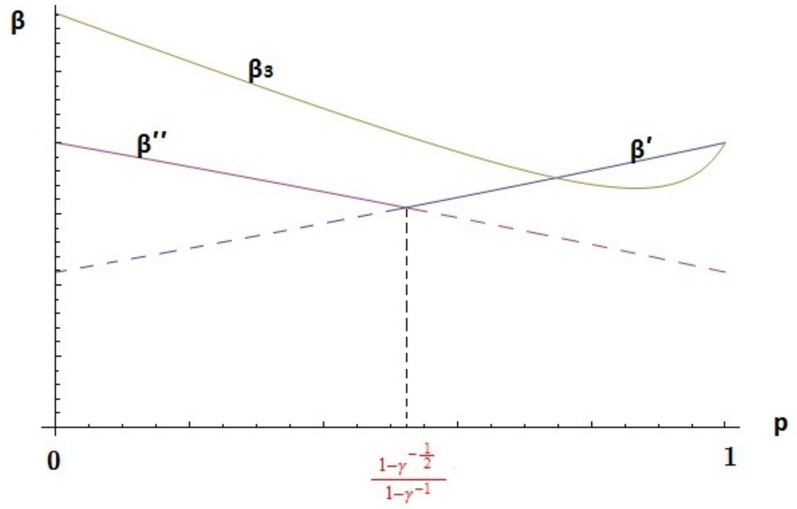
$$\begin{aligned}\beta_3 - \beta'' &= \frac{2(pq-1)[p(1-\gamma)+\gamma]}{[p(2q-1)-1]+(p-1)\gamma} - \frac{2[p(1-\gamma)+\gamma]}{p(1-\gamma)+2\gamma} \\ &= \frac{2(\gamma-1)[\gamma(1-p)+p](1-p)^2}{[\gamma(2-p)+p](1-p)(1+\gamma)}\end{aligned}$$

which are both positive under the above conditions. Thus we conclude that when PBE2 does not exist, $\beta_3 > \beta_1$.

Q.E.D.

A.4 Proof of Proposition 5

From the above proof we can see that when $q = 1$, $\beta_3 - \beta''$ is always positive. Then we can get the conclusion immediately.



Q.E.D.

A.5 Proof of Proposition 6

$$
\begin{aligned}
& W(PBE1) - W(PBE2) \\
&= \frac{qp}{\bar{c}} \int_0^{\beta\theta_B v} (\theta_B v - c)dc + \frac{1-q}{\bar{c}} \int_0^{\beta\theta_\emptyset v} (\theta_\emptyset v - c)dc + \frac{q(1-p)}{\bar{c}} \int_0^{\beta\theta_G v} (\theta_G v - c)dc \\
&- \frac{qp}{\bar{c}} \int_0^{\beta(\frac{1-q}{qp+1-q}\theta_\emptyset + \frac{qp}{qp+1-q}\theta_B)v} (\theta_B v - c)dc - \frac{1-q}{\bar{c}} \int_0^{\beta(\frac{1-q}{qp+1-q}\theta_\emptyset + \frac{qp}{qp+1-q}\theta_B)v} (\theta_\emptyset v - c)dc \\
&- \frac{q(1-p)}{\bar{c}} \int_0^{\beta\theta_G v} (\theta_G v - c)dc \\
&= \frac{1}{2}(2-\beta)\beta((1-q)\theta_\emptyset^2 + q\theta_G^2 + pq(\theta_B^2 - \theta_G^2))\frac{v^2}{\bar{c}} - \frac{(2-\beta)\beta}{2-2(1-p)q} \\
& (p^2q^2(\theta_B^2 - \theta_G^2) + pq(2(1-q)\theta_B\theta_\emptyset - (1-2q)\theta_G^2) + (1-q)((1-q)\theta_\emptyset^2 - q\theta_G^2))\frac{v^2}{\bar{c}}.
\end{aligned}
$$

By simplification, the above equation is eventually equal to $(2-\beta)\beta p(1-q)q(\theta_\emptyset - \theta_B)^2 v^2/(2-2(1-p)q)\overline{c}$ which is strictly larger than zero. Similarly, we can also get $W(PBE1) - W(PBE3)$ is strictly larger than zero.

Q.E.D.

A.6 Proof of Proposition 7

$$
\begin{aligned}
&W(PBE2) - W(PBE3) \\
&= \frac{qp}{\overline{c}} \int_0^{\beta(\frac{1-q}{qp+1-q}\theta_\emptyset + \frac{qp}{qp+1-q}\theta_B)v} (\theta_B v - c)dc + \frac{1-q}{\overline{c}} \int_0^{\beta(\frac{1-q}{qp+1-q}\theta_\emptyset + \frac{qp}{qp+1-q}\theta_B)v} (\theta_\emptyset v - c)dc \\
&+ \frac{q(1-p)}{\overline{c}} \int_0^{\beta\theta_G v} (\theta_G v - c)dc - \frac{qp}{\overline{c}} \int_0^{\beta\theta_B v} (\theta_B v - c)dc \\
&- \frac{1-q}{\overline{c}} \int_0^{\beta(\frac{1-q}{qp+1-q}\theta_G + \frac{qp}{qp+1-q}\theta_\emptyset)v} (\theta_\emptyset v - c)dc - \frac{q(1-p)}{\overline{c}} \int_0^{\beta(\frac{1-q}{qp+1-q}\theta_G + \frac{qp}{qp+1-q}\theta_\emptyset)v} (\theta_G v - c)dc.
\end{aligned}
$$

By substituting $\theta_\emptyset = p\theta_B + (1-p)\theta_G$ and simplifying, the above equation is equal to:

$$
-\frac{(2-\beta)\beta(1-q)qp}{2\overline{c}(1-q+(1-p)pq^2)}(1-3p+2p^2)(\theta_G - \theta_B)^2.
$$

The sign of above function is simply determined by $(1-3p+2p^2)$, which proves the proposition.

Q.E.D.

A.7 Proof of Proposition 8

(i) From the proof of Proposition 2, $\chi[r^*(\emptyset), \beta] = 0$ when $\beta = \beta' = 2\theta_B/(\theta_B + \theta_\emptyset)$ and $\Psi[r^*(\emptyset), r^*(G), \beta] = 0$ when $\beta = \beta'' = 2\theta_\emptyset/(\theta_\emptyset + \theta_G)$. Let $\theta_G/\theta_B \equiv \gamma$, we have:

$$
\frac{\partial \beta'}{\partial \gamma} = -\frac{2(1-p)}{[(1+p)+(1+p)\gamma]^2} < 0
$$

and

$$
\frac{\partial \beta''}{\partial \gamma} = -\frac{2p}{[p+(2-p)\gamma]^2} < 0.
$$

(ii) Similarly, as in the proof of proposition 2, the existence of PBE2 requires $1/2 < p < 1$ and $1 < \gamma < p(p-q+qp)/(1-p)^2(1-q)$. Under these conditions, we have:

$$
\frac{\partial \underline{\beta}_2}{\partial \gamma} = -\frac{2p[1-3(1-p)q+2(1-p)^2q^2]}{[p+\gamma(2-2q+p(2q-1))]^2} < 0
$$

and

$$\frac{\partial \overline{\beta}_2}{\partial \gamma} = -\frac{2(1-q+qp)(1-p)(1-q)}{[(1-q+p+qp)+(1-p)(1-q)\gamma]^2} < 0.$$

(iii) We have $\Psi[r^*(\emptyset), r^*(G), \beta_3] = 0$ where

$$\beta_3 = \frac{2(1-qp)\theta_\emptyset}{(1-qp)\theta_B + (1-q)\theta_\emptyset + (1-p)q\theta_G}.$$

The condition $0 < \beta_3 < 1$ implies that $1/2 < q < 1$ and $1/2q < p < 1$. Observe that:

$$\frac{\partial \beta_3}{\partial \gamma} = \frac{2(1-p)(1-qp)(1-2qp)}{[1+(1-p)\gamma+(1-2q)p]^2} < 0.$$

Q.E.D.

A.8 Proof of Proposition 9

(i) Since $\beta' = 2\theta_B/(\theta_B + \theta_\emptyset)$ and $\beta'' = 2\theta_\emptyset/(\theta_\emptyset + \theta_G)$, β' and β'' are independent of q.

(ii) We can verify the following:

$$\frac{\partial \underline{\beta}_2}{\partial q} = \frac{2(\gamma-1)[\gamma(1-p)+p](1-p)p}{[p+\gamma(2-2q+p(2q-1))]^2} > 0,$$

and

$$\frac{\partial \overline{\beta}_2}{\partial q} = \frac{2(\gamma-1)(1-p)p}{[1+p+\gamma(1-p)(1-q)-q+qp]^2} > 0.$$

(iii) It is straightforward to show:

$$\frac{\partial \beta_3}{\partial q} = -\frac{2(\gamma-1)[\gamma(1-p)+p](1-p)p}{[-1+\gamma(p-1)+p(2q-1)]^2} < 0.$$

Q.E.D.

A.9 Proof of Proposition 10

(i) Given that

$$\partial\beta'/\partial p = 2(\gamma-1)/((1+p)+(1-p)\gamma) > 0$$

and

$$\partial\beta''/\partial p = -2(\gamma-1)\gamma/[\gamma(2-p)+p]^2 < 0,$$

the relation between the existence of PBE1 and p depends on which of β' and β'' is higher. If, and only if, $\beta' < \beta''$, the increase in p will make it more likely for PBE1 to exist. Notice that

$$
\begin{aligned}
\beta' < \beta'' &\iff \frac{2}{(1+p)+(1-p)\gamma} < \frac{2p+2(1-p)\gamma}{p+(2-p)\gamma} \\
&\iff (\gamma-1)p^2 - 2\gamma p + \gamma > 0 \\
&\iff p < \frac{1-\gamma^{-\frac{1}{2}}}{1-\gamma} \ (\text{or } p > \frac{1-\gamma^{-\frac{1}{2}}}{1-\gamma}).
\end{aligned}
$$

Since $p \in (0,1)$ when and only when $p < (1-\gamma^{-\frac{1}{2}})/(1-\gamma)$, an increase in p will make it more likely for PBE1 to prevail.

(ii) We know that PBE2 exists when $0 < q < 1$ and $\frac{1}{2} < p < 1$ and

$$1 < \gamma < p(p-q+pq)/(1-p)^2(1-q).$$

Under this condition, we have that

$$\frac{\partial \underline{\beta}_2}{\partial p} = \frac{2(p^2q + \gamma(q-1)(2(1-p)^2-1) + \gamma((3-4p+p^2)q - 2(1-p)^2q^2-1))}{(p+\gamma(2-2q+p(2q-1)))^2} < 0,$$

and

$$\frac{\partial \overline{\beta}_2}{\partial q} = \frac{2(1-q)(\gamma-1)}{(1+p+\gamma(1-p)(1-q)-q+pq)^2} > 0.$$

(iii) For PBE3,

$$
\begin{aligned}
\frac{\partial \beta_3}{\partial p} &= -\frac{2(\gamma^2(p-1)^2q + 2pq + \gamma(q-1)(2p^2q-1) + p^2(q-2q^2)-1)}{[\gamma(p-1)+p(2q-1)-1]^2} \\
&= -\frac{2[(\gamma-1)q(2q+\gamma-1)p^2 + 2q(1-\gamma^2)p + (\gamma-1)(\gamma q+1)]}{[\gamma(p-1)+p(2q-1)-1]^2}
\end{aligned}
$$

So the sign is determined by the term $(\gamma-1)q(2q+\gamma-1)p^2+2q(1-\gamma^2)p+(\gamma-1)(\gamma q+1)$. As a function of p, the curve intersects with the horizontal axis at two points:

$$p = \frac{q+\gamma q \pm \sqrt{q-\gamma q-q^2+3\gamma q^2-2\gamma q^3}}{\gamma q - q + 2q^2}$$

It is easy to check that $p = (q+\gamma q+\sqrt{q-\gamma q-q^2+3\gamma q^2-2\gamma q^3})/(\gamma q-q+2q^2)$ is always larger than 1. So let $p^* \equiv (q+\gamma q-\sqrt{q-\gamma q-q^2+3\gamma q^2-2\gamma q^3})/(\gamma q-q+2q^2)$, we have that when $p < p^*$, β_3 is decreasing in p, and vice versa.
Q.E.D.

A.10 Proof of Proposition 11

Given the definition of welfare, we can directly obtain the following results.

$$\frac{\partial W(PBE1)}{\partial \beta} = (1-\beta)[(1-q)\theta_\emptyset^2 + (q-qp)\theta_G^2 + qp\theta_B^2] > 0,$$

$$\frac{\partial W(PBE2)}{\partial \beta} = \frac{1-\beta}{1-(1-p)q}[p^2q^2\theta_B^2 + 2pq(1-q)\theta_B\theta_\emptyset + (1-p)q[1-q(1-p)]\theta_G^2 > 0,$$

$$\frac{\partial W(PBE3)}{\partial \beta} = \frac{(2-\beta)\beta}{2(1-pq)}[(pq-p^2q^2)\theta_B^2+(1-q)^2\theta_\emptyset^2+q^2(1-p)^2\theta_G^2+2q[(1-q)(1-p)]\theta_\emptyset\theta_G] > 0.$$

It is straightforward to show that

$$\frac{\partial W(PBE1)}{\partial q} = \frac{1}{2}(2-\beta)\beta(1-p)p(\theta_G-\theta_B)^2V^2 > 0,$$

$$\frac{\partial W(PBE2)}{\partial q} = \frac{(2-\beta)\beta(1-p)p^2(\theta_G-\theta_B)^2V^2}{2[1+(p-1)q]^2} > 0,$$

$$\frac{\partial W(PBE3)}{\partial q} = \frac{(2-\beta)\beta(1-p)^2p(\theta_G-\theta_B)^2V^2}{2(1-pq)^2} > 0.$$

It is also straightforward to show that

$$\frac{\partial W(PBE1)}{\partial p} = \frac{1}{2}(2-\beta)\beta(\theta_G-\theta_B)[(2p-2pq+q-2)\theta_G - (2p-2pq+q)\theta_B]V^2 < 0,$$

$$\frac{\partial W(PBE2)}{\partial p} = -\frac{(2-\beta)\beta(\theta_G-\theta_B)}{[1+(p-1)q]^2}[2(1-q)^2\theta_G+(2p(1-q)+p^2q)[\theta_B+(2q-1)\theta_G]] < 0,$$

$$\frac{\partial W(PBE3)}{\partial p} = -\frac{(2-\beta)\beta(\theta_G-\theta_B)}{2(1-pq)^2}[q(\theta_B-\theta_G)+2\theta_G+(p^2q-2p)[(2q-1)\theta_B+\theta_G]]V^2 < 0.$$

Q.E.D.

A.11 Proof of Existence of Equilibria for the Extended Model

Following the analysis of Proposition 2, we can obtain 16 PBE's in the absence of refinement by the intuitive criterion. The pattern of these 16 PBE's resembles Table 1 except for PBE3h in which we have amnesia. After applying the intuitive criterion, we can identify the 16 possible PBE's listed in Table 1. Let the reliability of signal $\emptyset$ and signal G to self-0^+ be $r^*(\emptyset)$ and $r^*(G)$ respectively, and that to self-1 be $r^{**}(\emptyset)$ and $r^{**}(G)$.

Take PBE3c as an example. This equilibrium is determined by the off-equilibrium belief $r^*(\emptyset)$. In this equilibrium, no matter what $r^*(\emptyset)$ is, delusion is always strictly better than correct recall for self-0 of type-$\emptyset$; while for self-0 of type-B, correct recall and amnesia are indifferent when $r^*(\emptyset) = 0$. After the refinement, when self-1 receives an empty signal, he will know that he is type-B, and thus the amnesia becomes fake.

PBE3e offers another example. Here, self-0^+ of type-$\emptyset$ strictly prefers delusion rather than correct recall, while self-0^+ of type-B is indifferent between correct recall and amnesia when $r^{**}(\emptyset) = 0$. Thus, after refinement, the belief can only be $r^{**}(\emptyset) = 0$.

We can similarly refine the pre-intuitive-criterion equilibria corresponding to PBE3a, PBE3f, PBE3h, PBE4a and PBE4b and obtain the results summarized in Table 1.

Q.E.D.