NANYANG TECHNOLOGICAL UNIVERSITY

Division of Economics, EGC
School of Humanities and Social Sciences
Nanyang Technological University
14 Nanyang Drive
Singapore 637332

Sunk Cost as a Self-Disciplining Device

Fuhai HONG and Xiaojian ZHAO

18 August 2015

EGC Report No: 2015/03

HSS-04-88
Tel: +65 67905689
Email: D-EGC@ntu.edu.sg
http://egc.hss.ntu.edu.sg

Sunk Cost as a Self-Disciplining Device*

Fuhai Hong†   Xiaojian Zhao‡

November 2014

Abstract

Building on an intra-personal self-signaling game, the paper provides an economic model to show that the sunk cost effect may stem from an attempt to overcome the under-investment problem associated with present bias. The current self may take a costly action (which is a sunk cost for the future self) to signal the individual's ability that motivates his future self-disciplining behaviors. In equilibrium, a higher level of sunk cost gives rise to a higher probability for the individual to continue the project.

Keywords: Present bias, signaling, sunk cost fallacy, limited memory.
JEL Classification: D03; D83; Z13

*We thank Teh Tat How for excellent research assistance, and Alexander White, Huaxiang Yin for helpful comments.

†Division of Economics, Nanyang Technological University, Singapore (fhhong@ntu.edu.sg).

‡Department of Economics, Hong Kong University of Science and Technology, Hong Kong (xjzhao@ust.hk).

"If someone offered us a pill that henceforth would make us people who never honored sunk costs, we might be ill advised to accept it ..."

Robert Nozick, *The Nature of Rationality* (1993, p. 23).

1 Introduction

Why do people, especially gangsters, make painful tattoos?[1] Life in a gang is tough. Yes, the gangsters may want to signal to others that they are also tough enough to deter potential enemies. But doesn't that also remind themselves of a sort of determination? Why do people in a relationship write love letters and decorate costly pre-wedding and wedding photos? These may be a signal to the partner on how he loves her. But aren't the photos and letter drafts also indicative of something to oneself?

Standard economics textbooks teach us that we should never cry over spilled milk because sunk costs are sunk. Nevertheless, consider a person who is contemplating whether to terminate a relationship. When looking at all the letter drafts, decorated photos, and the tickets for shows they watched together and parks they visited together, he hesitates. Would you say this is a fallacy because all of these are sunk costs? Consider a gangster contemplating about quitting the gang. He hesitates as he sees his tattoos on the skin. Does he commit a sunk cost fallacy?

Present bias or the tendency to pursue immediate gratification is a prevalent psychological trait in inter-temporal decision-making, and could lead to inefficient under-investment in long-run projects (see, e.g. O'Donoghue and Rabin, 1999). An individual knows that doing exercise is good for health but the cost of doing so is immediate (and the temptation of the comfort of staying at home is hard to resist) while the reward is only realized in the future. The immediate cost seems so salient that the individual may thus delay doing exercise to tomorrow, while when tomorrow comes he delays again. On the other hand, the individual may have bought a gym membership and when he struggles about whether to go to the gym, another behavioral bias may arise: Maybe I should go to the gym, since I have paid so much money to buy the membership! Therefore, the familiar sunk cost effect arises. Interestingly,

[1] For example, Japan's "yakuza" organization the Yamaguchi-gumi, which is the largest criminal organizations in the world (see *Fortune 5: The biggest organized crime groups in the world,* Fortune Magazine, 14 September 2014), has members with almost full-body tattoos (see *Japan godfather released from jail,* Mafia Today, 10 April 2011).

the sunk cost effect works in the opposite direction to the effect of present bias, and alleviates the under-investment problem in this example.

To the best of our knowledge, the American philosopher Robert Nozick (1993) is the first to discover the instrumental role of the sunk cost effect in resisting temptations.[2] While standard economics textbooks regard the sunk cost effect as a fallacy, there are several studies in economics attempting to provide rationales for this effect.[3] Among them, Eyster (2002) explains it as the result of preferences over rationalizing past choices. Instead of attributing the fallacy to the affective reason, other models adopt the cognitive approach, and find the inter-personal signaling value of the sunk cost effect. For example, Kanodia, Bushman and Dickhaut (1989) and McAfee, Mialon and Mialon (2010) study the role of honoring sunk cost in building reputations.

More recently, Baliga and Ely (2011) provide a rationale for the sunk cost fallacy in a two-period intra-personal model with time consistent preferences. In the first period, an individual decides on whether to initiate a project with some initiation cost; in the second period, if he has initiated the project, the individual decides on whether to complete it with some continuation cost. The individual knows the expected value of the project when initiating the project but he forgets it in the second period. Therefore, the initiation cost, which is sunk from the viewpoint at the second period, is informative when the individual decides on whether to complete the project. A higher sunk cost indicates higher expected value and makes completion more likely, which they call a Concorde effect.[4]

In Baliga and Ely's (2011) model, initiation is an endogenous decision; the individual learns the expected value of the project before initiating the project, and no further information comes in between initiation and completion, except the continuation cost. However, for most if not all long-run projects, people learn new information about the type of the project, after they have initiated a project, and such information is no less important than the information collected before the initiation. For examples, a rookie PhD student typically learns how the academia runs only after he started his PhD study, rather than when he contemplated whether to start a PhD or not; a wife knows her husband better than when the husband just became her new

[2]Nozick (1993, p. 23) writes, "We can knowingly employ our tendency to take sunk costs seriously as a means of increasing our future rewards. If this tendency is irrational, it can be rationally utilized to check and overcome another irrationality."

[3]For psychological studies on this effect, see, e.g. Arks and Blumer (1985).

[4]They also consider a different version of sunk cost fallacy, where the sunk cost signals the future continuation cost, rather than the expected value. In this model, a higher sunk cost discourages further investment, which they call a *pro-rata effect*.

boyfriend. In these cases, the sunk initiation cost is uninformative about the information learnt after the project initiation.

In line with the idea of Nozick (1993), our paper provides an alternative economic model with an endogenous sunk cost effect. We are different from Baliga and Ely (2011) in two main aspects. First, we will focus on the information learnt after the project initiation. For simplicity and to clearly distinguish our model from Baliga and Ely (2011), we assume exogenous initiation of the project. Second and more importantly, we introduce present-biased preference and investigate the interaction between present bias and limited memory. More specifically, we consider the following three period model. In the first period, an individual learns the type of his project, and then takes a costly action. As in Baliga and Ely (2011), the individual has limited memory. He remembers the costly action but forgets his type in the second period when he has to determine whether to complete the project. If completed, the project yield returns in period 3. Our results are as follows. First, in a benchmark case where the individual's preference is time consistent, the individual does not invest in the costly action, and therefore there is no a sunk cost effect. This is in contrast with Baliga and Ely (2011). Second, when the individual has present bias, there is a refined separating equilibrium in which the individual invests in the costly action; moreover, the sunk cost effect emerges, in that the action taken in period 1, which is sunk from the viewpoint of period 2, influences the completion decision in period 2, although it affects neither the return of the project nor the cost of completion.[5] Particularly striking is that with a continuum of types we show that the probability of continuing the project in period 2 is strictly increasing in the level of sunk cost in period 1. While Nozick (1993) treats the sunk cost effect as exogenous in discussing its value in resisting temptations, the sunk cost effect is endogenous in our model. In a nutshell, Baliga and Ely (2011) build a model with time consistent preference where the sunk initiation cost has informational value about the project return, while our model with conflict of interest between multiple selves is in essence a classic signaling story where the current self has an incentive to burn money as a costly signaling device so as to discipline the future self, since the money burned as sunk cost will be taken into account by the future self in decision making.

Like in our model, Bénabou and Tirole (2004) and Ericson (2014) consider the interplay between limited memory and present bias. Bénabou and Tirole

[5] Our result is consistent with the observation that only intentionally made sunk costs, but not accidentally made sunk costs, will be considered when making the continuation/completion decision (see Doody, 2013).

(2004) analyze how individuals develop personal rules such as diets, smoking only after meals, monthly saving targets, to deal with dynamic inconsistency. In their model, at each period, the individual decides on whether to initiate a willpower activity (like abstinence in drinking, smoking or spending), and by doing so, he puts his will to test, because he will have to decide on whether to give up or persevere under present biased preference later in that period. It is shown that exercising willpower motivates the future individual not to give up given his imperfect recall. In Ericson (2014), memory is limited in the sense that an individual may forget to complete a task in the future, rather than some event occurring before. Ericson shows that limited memory can function as a commitment device for present-biased individuals, as the anticipation of forgetting to finish the task in the future may alleviate the problem of procrastination. Notably, neither Bénabou and Tirole (2004) nor Ericson (2014) analyzes the sunk cost effect though.

The rest of this paper is organized as follows. The next section presents the model basics. Section 3 analyzes the case with two types while Section 4 is devoted to continuous types. The final section concludes. The proofs of all propositions are relegated to the appendix.

2 Model Basics

Our model builds along the line of the canonical self-management games by Carrillo and Mariotti (2000) and Bénabou and Tirole (2002). There are three dates, $t = 0, 1, 2$. We allow for present-biased preferences by assuming that the individual at t discounts expected payoffs at $t + n(> t)$ with a discount factor equal to $\beta\delta^n$, where δ normalized to 1 is the normal discounting factor, and $\beta \in (0, 1)$ corresponds to hyperbolic discounting.[6]

The timing is as follows.

At $t = 0$, the individual or self-0 has started a project,[7] and privately obtains the information about his ability, or the project's probability of success, θ. Then he chooses tangible and observable $a \in \mathbb{R}^+$ to signal his ability to his future self-1, while a does not affect θ.[8] The cost of a is $c(\theta, a)$, which captures the instantaneous subjective disutility (either physical or psychological cost) of making a and will be unobservable from self-1's viewpoint at $t = 1$.

[6]See the seminal works by Strotz (1955) and Laibson (1997).

[7]In this respect, our model is different from Baliga and Ely (2011) in which the starting decision is endogenous, and the individual infers the project's value from this decision.

[8]Although our assumption of money burning is a simplifying one, capturing "sink(ing) resources and efforts" in terms of Nozick (1993, p. 23), like Spence's (1973) education-as-a-signal model, it can beem relaxed to allow a productive action without qualitatively changing the insights.

Assume that c is smooth, that $c_2(\theta, a) > 0$, $c_{22}(\theta, a) > 0$, and $c_{12}(\theta, a) < 0$ for all θ, a, and that $c(\theta, 0)$ is normalized to 0 for all θ. Thus, we impose standard increasingness and convexity of the cost function. Moreover, there is a low marginal cost for high-ability individuals. In the gangster example, a is the tattoo on the skin and c is the magnitude of pain incurred from making the tattoo; in the dating example, a is the love letter drafts and pre-wedding and wedding photos, while c is the time and effort spent on the letters and photos.

At $t = 1$, the individual or self-1, who does not know θ but could infer it upon observing a in the last period, decides whether to continue the project at an add-on cost k, which is unknown by self-0, but self-0 knows k's continuous density function f distributed over $\mathbb{R}^+$. As in Bénabou and Tirole (2004) and Baliga and Ely (2011), while the individual may forget his motivations, it may be easier to remember his observable actions in the past, which generate sunk costs.

At $t = 2$, the project outcome is realized. The project will yield benefit $V > 0$ if it is continued at $t = 1$ and successful at $t = 2$ and zero otherwise.

In the case of complete information, self-1 chooses to continue if and only if $\beta\theta V \geq k$. However, from self-0's point of view (a.k.a. the long-run perspective as in O'Donoghue and Rabin (1999)), self-1 under-invests as he should have invested as long as $\theta V \geq k$. Self-0 of type-θ would choose a to maximize

$$-c(\theta, a) + \beta \int_0^{\beta\theta V} (\theta V - k) f(k) dk$$

with the solution $a = 0$ for all θ.

However, in the case of incomplete information as we assume here, self-1's decision depends on his belief about θ. Since self-1 infers $\widehat{\theta}$ from self-0's a, the optimal a chosen by self-0 may not be zero as in the case of complete information. For the analysis, we adopt the solution concept of perfect Bayesian equilibrium (PBE) in pure strategies (with certain refinements as we show below).[9]

3 Two Types

In this section, we assume that θ can take only two values, and let $\theta \in \{\theta_L, \theta_H\}$ with $0 < \theta_L < \theta_H < 1$. θ_H occurs with probability p and θ_L occurs with probability $1 - p$. Let $\overline{\theta} = p\theta_H + (1 - p)\theta_L$. Let $\mu(a)$ be the belief of

[9] Mailath (1992) and Mailath et al (1993) justify the focus on pure strategies only and show that mixed strategies are unimportant for signalling games.

self-1 on the probability of type-H conditional on a. Let $\widehat{\theta}(a) = \mu(a)\theta_H + (1-\mu(a))\theta_L$ be the posterior expected type for self-1.

3.1 Benchmark: Time Consistent Preference

With time consistent preference, there is no conflict of interest between self-0 and self-1, leading to neither type to want to mimic the other. In this case, we have the following observation.

Observation 1 *In the case of time consistent preference,*

a) there exists no PBE surviving the Intuitive Criterion (Cho and Kreps, 1987) in which some type of the individual chooses a positive a;

b) if, and only if, $\Gamma \equiv \int_{\bar{\theta}V}^{\theta_H V} (\theta_H V - k) f(k)dk < \int_{\theta_L V}^{\bar{\theta}V} (k - \theta_L V) f(k)dk \equiv \Delta$, and there exists a' such that $c(\theta_H, a') = \Gamma < c(\theta_L, a') = \Delta$, there is a pooling equilibrium where individuals of both types choose $a = 0$ surviving the intuitive criterion.

The observation suggests that no individual will choose a positive a. With time consistent preference, self-0 does not invest in a in equilibrium, with the refinement of the Intuitive Criterion.

The sunk cost effect is therefore not relevant in this case. This is in contrast with Baliga and Ely (2011) who find the sunk cost effect in the presence of time consistent preference. In their model, the cost of initiating a project is exogenously given as a random variable; after a project has been initiated, this sunk cost signals its expected value. However, in our model, a is a choice variable, leading to zero investment in it under time consistent preference.

Notably, with time consistent preference, self-0 and self-1 have common interest. Thus, back to the case with continuous a, although self-0 does not invest in a to signal his type, self-0 can "talk" to self-1 about his type, and self-1 has no incentive not to trust self-0.

It is worth mentioning that we have assumed a continuous variable a. However, if we assume instead that there exists a smallest positive value of a, denoted by ϵ, which is arbitrarily close to 0, then there exist separating equilibria in which $a_H = \epsilon > a_L = 0$ or $a_L = \epsilon > a_H = 0$. Moreover, they are qualified as a unique set of the undefeated separating equilibria (Mailath et al, 1993) surviving the Intuitive Criterion. Appendix B illustrates the details of the analysis. Still, with a smallest positive a, self-0 does not invest in a non-trivially.

3.2 Present-Biased Preference

We now consider the non-trivial case of time inconsistent preference. To make the analysis have a sharp focus, we assume that $\beta \leq \theta_L/\theta_H$. That is, the problem of present bias is sufficiently severe.

A self-0 $i \in \{\theta_L, \theta_H\}$ chooses a to maximize

$$-c(\theta_i, a) + \beta \int_0^{\beta\widehat{\theta}V} (\theta_i V - k)\, f(k)dk.$$

Since $\beta \leq \theta_L/\theta_H$, we have $\beta\widehat{\theta}V \leq \theta_L V < \theta_H V$ for any $\widehat{\theta}$; it is thus impossible for some type of the individual to have over-investment. Self-1 of both types (weakly) under-invests and the payoff of self-0 of any type is increasing in $\widehat{\theta}$. In particular, self-0 of type-L wants to pretend to be type-H. We first examine the pooling equilibrium in which self-0 of both types chooses the same a, and self-1 cannot tell the type.

Pooling Equilibrium The result of pooling equilibrium is summarized in the following proposition.

Proposition 1 *With two types of individuals,*

a) there exist pooling equilibria where individuals of both types choose $a^p > 0$ if a^p is not too large;

b) no pooling equilibrium survives the intuitive criterion.

Note that the result of pooling equilibrium in Proposition 1-a) is akin to the result of building self-reputation via self-restraint in Bénabou and Tirole (2004). In their model, both types of individuals have the option of attempting self-control in order to signal to future selves not to give up. Pooling equilibrium will occur in which the low type can mimic the high type by exercising will-power. However, Proposition 1-b) is in contrast to Bénabou and Tirole (2004). Since their model incurs a binary choice of attempting self-control, whereas our model involves a continuous variable a without any upper bound, the pooling result in Bénabou and Tirole (2004) survives the intuitive criterion.

In addition, the pooling equilibrium cannot explain the sunk cost effect in which a higher probability of continuing the project is associated with a higher a. Thus, we focus on the analysis on separating equilibrium.

Separating Equilibrium The result of separating equilibrium is summarized in the following proposition.

Proposition 2 *With two types of individuals, there is a unique separating equilibrium surviving the intuitive criterion in which self-0 of type-H chooses* $a_{s,L} > 0$ *and self-0 of type-L chooses 0 where* $a_{s,L}$ *is given by (2) in the appendix.*

Thus, although self-0 with time consistent preference does not invest in a, self-0 with present-biased preference invests in a nontrivially. $a_{s,L}$ is the lowest investment level that makes self-0 of type-L has no incentive to mimic type-H.[10] Therefore, the equilibrium in Proposition 2 represents a least-cost separating equilibrium.

With severe present bias, self-0 of type-L always has an incentive to mimic type-H to overcome the under-investment problem associated with time-inconsistent preference. Therefore, self-0 of type-H invests in a positive amount a to signal its own type. Self-1 will take into account a when making the continuation decision. In the separating equilibrium, since $\widehat{\theta}(a_{s,L}) = \theta_H > \widehat{\theta}(0) = \theta_L$, the positive, sunk $a_{s,L}$ signals the high type, giving rise to a higher probability of continuation compared to the observation of no investment in a. This result smacks of the sunk cost effect. As Nozick (1993, p. 22) argues, "we do not treat the past effort we have devoted to ongoing projects of work or life as of no account ... Such projects help to define our sense of ourselves and of our lives." Our results thus explain the painful tattooing and the hesitation about quitting when seeing tattoos on the skin for the gangster example discussed in the Introduction. A similar explanation applies to the dating example.

However, given that there are only two types, we cannot check the monotonicity of the probability of continuation with respect to a. Therefore, we consider a continuum of types in the following section.

4 A Continuum of Types

In this section, we assume a continuum of types θ, which is continuously distributed over $[\underline{\theta}, \overline{\theta}]$ where $0 < \underline{\theta} < \overline{\theta} < 1$. As in the two-type case, similarly, we assume that $\beta < \underline{\theta}/\overline{\theta}$.

In this case, self-0 of each type θ chooses a, from which self-1 draws inference about θ, and reacts accordingly. Self-0's expected payoff is given

[10] In the case where β is large enough such that $a_{s,L}$ does not exist, implying that self-0 of type-L has no incentive to mimic self-0 of type-H, the problem is essentially the same as in the case of time consistent preference, i.e. $\beta = 1$.

by

$$U(\theta, \widehat{\theta}, a) = -c(\theta, a) + \beta \int_0^{\beta \widehat{\theta}(a) V} (\theta V - k) f(k) dk. \tag{1}$$

Self-0 thus chooses a mapping $A : [\underline{\theta}, \overline{\theta}] \to \mathbb{R}^+$ representing a choice a for each type θ, taking into account self-1's inference $\widehat{\theta}(a)$. In separating equilibrium, A is one-to-one and incentive compatible. Thus, self-0 of type-θ chooses a to maximize the objective in (1). The following proposition presents the least-cost separating equilibrium surviving the intuitive criterion.

Proposition 3 *With a continuum of types, A is differentiable everywhere, and the following differential equation (DE) and boundary condition (BC) characterize self-0's incentive compatible separating solution:*

$$A'(\theta) = \frac{(1-\beta)\,\beta\theta V^2 f(\beta\theta V)}{c_2(\theta, a)}, \tag{DE}$$

and

$$A(\underline{\theta}) = 0. \tag{BC}$$

By (DE), note that $A'(\theta) > 0$. The higher the individual's ability θ, the larger the amount of money a burned by self-0, which gives rise to a more confident belief about θ held by self-1. Note that a higher confidence level $\widehat{\theta}$ leads to a greater cutoff value of the add-on cost k, making the continuation of the project more likely to occur at $t = 1$. Thus, we have the following corollary.

Corollary 1 *For self-1, the probability of continuing the project is increasing in the observed a chosen by self-0.*

Therefore, we rationalize the sunk cost effect where a high level of a, which is sunk for self-1, entails a large incentive for the individual to continue the project, as a result of the signal of the individual's high ability θ.

Notably, Baliga and Ely (2011) discuss two cases on whether the exogenous initiation cost and the expected value of the project are correlated or not. Our result suggests that even if a cannot influence θV directly, a and θV are correlated in self-1's belief in our least-cost separating equilibrium.

5 Conclusion

For an economist teaching introductory economics, one of the key elements to educate undergraduate students is the notion of sunk cost. Particularly insightful is that, when an individual facing an incremental cost decides whether or not to continue a project, a rational individual should not take the level of sunk cost into account (let the past be past!). However, a large evidence in economics and psychology suggests that it is not the case when it comes to the sunk cost effect in reality (see, e.g., Arks and Blumer, 1985).

While the behavioral industrial organization literature has started to examine how consumers' prevalent sunk cost effect may affect optimal two-part pricing (e.g. Wang and Yang, 2010), our paper provides a different angle by studying why the sunk cost effect emerges.[11] Several papers in economic theory tend to provide rationales for the sunk cost effect, including from the perspective of social-signaling (as reviewed in the Introduction). In our paper, we focus on self-signaling rather than social-signaling. Closest to our paper is the model of memory kludge by Baliga and Ely (2011). Both models are information-based, and rely on the level of sunk cost to infer the expected value of the project to make the correct decision. However, Baliga and Ely (2011) study two selves without any conflict of interests. Facing an exogenous cost, the first self decides whether to start the project. Then the second self infers the project's value from this starting decision as well as the exogenous cost observable. In our paper, we focus on the information learnt after the project starts. The sunk cost effect arises only under time inconsistent preference where there is a conflict of interest between the two selves: Self-0 takes a costly action to signal the value of the project in order to motivate present-biased self-1 to continue the project.

Our paper belongs to a more general agenda in economics and psychology that has attempted to rationalize the observed cognitive biases. Stable natural and social environments generate certain cognitive patterns, and their induced behaviors. The cognitive and behavioral patterns persist when they have an adaptive role. Carrillo and Mariotti (2000) and Bénabou and Tirole (2002, 2004) have investigated the instrumental value of ignorance, overconfidence and personal rules with "escalation of commitment" for individuals with time-inconsistent preferences. By adopting a similar set-up, Dessi and Zhao (2013) study the well-documented differences between East-Asians and Westerners in terms of self-esteem and social emotions, and their motivational values. Chew, Huang and Zhao (2014) check a variety of memory biases to supply over-confidence as a motivational device. All these mod-

[11] While the sunk cost effect serves as a motivational device in the investment problem in our model, this effect may not be functional in some cases. In some circumstance, it may reinforce the adverse effect due to present bias. For example, Just and Wansink (2011) find that the flat-rate pricing such as "all-you-can-eat" buffet pricing can take advantage of the sunk cost effect and thus exacerbates the over-consumption tendency associated with obesity.

els and our paper take the time-inconsistent preferences and the associated under-investment problem as given, and investigate the functional role of certain cognitive biases, which emerge and are sustained in equilibrium.[12]

Fudenberg (2006) argues that "... behavioral economists (and economic theorists!) should devote more effort to synthesizing existing models and developing more general ones, and less effort to modeling yet another particular behavioral observation." In our paper, the self-signaling model revisits the classic sunk cost effect, and relates it to present-biased preference based on the classic approach of information economics. Our theory thus provides a rationale for the sunk cost effect. In the future, more works can be done on investigating how different psychological traits could emerge endogenously from various intrapersonal or interpersonal interactions.

[12] Different from assuming time inconsistent preference, alternative approaches on rationalizing behavioral biases such as collective delusions and risk attitudes include more recent works with exogenous preference over beliefs by Bénabou (2013) and Gottlieb (2014).

References

[1] Arkes, H. and C. Blumer (1985). "The Psychology of Sunk Cost," *Organizational Behavior and Human Decision Processes*, 35, 1224-140.

[2] Baliga, S. and J. Ely (2011). "Mnemonomics: The Sunk Cost Fallacy as a Memory Kludge," *American Economic Journal: Microeconomics*, 3, 35-67.

[3] Bénabou, R. (2013). "Groupthink: Collective Delusions in Organizations and Markets," *Review of Economic Studies*, 80, 429-462.

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

[5] Bénabou, R. and J. Tirole (2004). "Willpower and Personal Rules," *Journal of Political Economy,* 122(4), 848-886.

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

[7] Chew, S., W. Huang and X. Zhao (2014). "Selective Memory and Motivated Delusion: Theory and Experiment," mimeo, Hong Kong University of Science and Technology.

[8] Cho, I., and D. Kreps (1987). "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics,* 102, 179-221.

[9] Dessi, R. and X. Zhao (2013). "Self-Esteem, Shame and Personal Motivation," IDEI Working Paper, n. 639.

[10] Doody, R. (2013). "The Sunk Cost 'Fallacy' Is Not a Fallacy," mimeo, Department of Linguistics and Philosophy, MIT.

[11] Ericson, K. (2014). "On the Interaction of Memory and Procrastination: Implications for Reminders," NBER Working Paper 20381.

[12] Eyster, E. (2002). "Rationalizing the Past: A Taste for Consistency," mimeo, Nuffield College, Oxford.

[13] Fudenberg, D. (2006). "Advancing Beyond 'Advances in Behavioral Economics,'" *Journal of Economic Literature*, 694-711.

[14] Gottlieb, D. (2014). "Imperfect Memory and Choice under Risk," *Games and Economic Behavior*, 85, 127-158.

[15] Just, D.R. and B. Wansink (2011). "The Flat-Rate Pricing Paradox: Conflicting Effects of "All-You-Can-Eat" Buffet Pricing," *Review of Economics and Statistics*, 93(1), 193-200.

[16] Kanodia, C. and R. Bushman and J. Dickhaut (1989). "Escalation Errors and the Sunk Cost Effect: An Explanation Based on Reputation and Information Asymmetries, " *Journal of Accounting Research*, 27(1), 59-77.

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

[18] Mailath, G.J. (1987). "Incentive Compatibility in Signaling Games with a Continuum of Types," *Econometrica*, 55(6), 1349-1365.

[19] Mailath, G.J. (1992). "Signalling Games," in *Recent Development in Game Theory* (J. Creedy, J. Borland and J. Eichberger, Eds.), Edward Elgar Publishing Limited, Hants, England.

[20] Mailath, G.J., M. Okuno-Fujiwara, and A. Postlewaite (1993). "Belief-Based Refinements in Signalling Games," *Journal of Economic Theory*, 60, 241-276.

[21] Mailath, G.J., and E. von Thadden (2013). "Incentive Compatibility and Differentiability: New results and classic applications," *Journal of Economic Theory*, 148, 1841-1861.

[22] McAfee, R.P, H.M. Mialon and S. H. Mialon (2010). "Do Sunk Costs Matter?" *Economic Inquiry*, 48(2), 323-36.

[23] Nozick, R. (1993) *The Nature of Rationality*, Princeton University Press, New Jersey.

[24] O'Donoghue, T. and M. Rabin (1999). "Doing It Now or Later," *American Economic Review*, 89, 103–24.

[25] Spence, M. (1973). "Job Market Signaling," *Quarterly Journal of Economics*, 87(3), 355-374.

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

[27] Wang, X. and B. Yang (2010). "The Sunk-Cost Effect and Optimal Two-Part Pricing," *Journal of Economics*, 101, 133-148.

A Proof of Observation 1

a) With pure strategies, a PBE is either a pooling or separating equilibrium. We proceed the proof in two steps.

Step 1. There is no pooling equilibrium in which both types of individuals choose a positive a.

Suppose there is a pooling equilibrium in which both types of individuals choose $a_P > 0$. Then $\mu(a_P) = p$. Consider a small positive number ε. If $\mu(a_P - \varepsilon) \leq p$, then $\theta_L \leq \widehat{\theta}(a_P - \varepsilon) \leq \overline{\theta}$. Below shows that self-0 of type-L has incentive to deviate:

$$
\begin{aligned}
& -c(\theta_L, a_P) + \int_0^{\overline{\theta}V} (\theta_L V - k) f(k) dk \\
& = -c(\theta_L, a_P) + \int_0^{\widehat{\theta}(a_P-\varepsilon)V} (\theta_L V - k) f(k) dk \\
& + \int_{\widehat{\theta}(a_P-\varepsilon)V}^{\overline{\theta}V} (\theta_L V - k) f(k) dk \\
& \leq -c(\theta_L, a_P) + \int_0^{\widehat{\theta}(a_P-\varepsilon)V} (\theta_L V - k) f(k) dk \\
& < -c(\theta_L, a_P - \varepsilon) + \int_0^{\widehat{\theta}(a_P-\varepsilon)V} (\theta_L V - k) f(k) dk.
\end{aligned}
$$

where the first inequality comes from $\int_{\widehat{\theta}(a_P-\varepsilon)}^{\overline{\theta}V} (\theta_L V - k) f(k) dk \leq 0$ because of $\theta_L V \leq \widehat{\theta}(a_P - \varepsilon) V$. Similarly, we can show that if $\mu(a_P - \varepsilon) \geq p$, then self-0 of type-$H$ has incentive to deviate. Therefore there is no pooling equilibrium in which both types of individuals choose a positive a.

Step 2. There does not exist a separating equilibrium surviving the Intuitive Criterion.

Recall that with time-consistent preference, each type of self-0 has no incentive to mimic the other type. Suppose that there is a separating equilibrium where $a_H > a_L$. Then for any $a' \in (a_L, a_H)$,

$$
-c(\theta_L, a_L) + \int_0^{\theta_L V} (\theta_L V - k) f(k) dk > -c(\theta_L, a') + \int_0^{\theta_L V} (\theta_L V - k) f(k) dk
$$

and

$$
-c(\theta_H, a_H) + \int_0^{\theta_H V} (\theta_H V - k) f(k) dk < -c(\theta_H, a') + \int_0^{\theta_H V} (\theta_H V - k) f(k) dk.
$$

Therefore, observing an off-equilibrium action a', self-1 would infer that the action is taken by self-0 of type-H. Thus self-0 of type-H has incentive to deviate to a'. Hence, the equilibrium, if exists, does not survive the Intuitive Criterion. Similarly, we can show that a separating equilibrium where $a_L > a_H$, if exists, does not survive the Intuitive Criterion.

b) We proceed the proof in the following steps.

Step 1. There is a pooling equilibrium in which both types of individuals choose $a_P = 0$.

Such an equilibrium exists, if we let the belief system be $\mu(a) = p$ for any a, consistent with equilibrium strategies. Then no type of the individuals has incentive to deviate to a more costly $a > 0$.

Step 2. If and only if $\Gamma < \Delta$, and there exists a' such that $c(\theta_H, a') = \Gamma < c(\theta_L, a') = \Delta$, the pooling equilibrium with $a_P = 0$ survives the Intuitive Criterion.

To show it, by the definition of the Intuitive Criterion, if and only if there exists a' such that either

$$\begin{cases} \int_0^{\overline{\theta}V} (\theta_L V - k) f(k)dk > -c(\theta_L, a') + \int_0^{\theta_L V} (\theta_L V - k) f(k)dk \Leftrightarrow c(\theta_L, a') > \Delta \\ \int_0^{\overline{\theta}V} (\theta_H V - k) f(k)dk < -c(\theta_H, a') + \int_0^{\theta_H V} (\theta_H V - k) f(k)dk \Leftrightarrow c(\theta_H, a') < \Gamma \end{cases},$$

or,

$$\begin{cases} c(\theta_L, a') < \Delta \\ c(\theta_H, a') > \Gamma \end{cases},$$

then the equilibrium with $a_P = 0$ does not survive the Intuitive Criterion. Suppose $\Gamma < \Delta$. Unless there exists a' such that $c(\theta_H, a') = \Gamma$ and $c(\theta_L, a') = \Delta$, we can always find an a'' such that either $\Gamma < c(\theta_H, a'') < c(\theta_L, a'') < \Delta$ or $c(\theta_H, a'') < \Gamma < \Delta < c(\theta_L, a'')$. Suppose $\Delta \leq \Gamma$. Then there always exists a' satisfying $c(\theta_L, a') > \Delta$ and $c(\theta_H, a') < \Gamma$. Therefore, except in a special case where there exists a' such that $c(\theta_H, a') = \Gamma < c(\theta_L, a') = \Delta$, the pooling equilibrium with $a_P = 0$ does not survive the Intuitive Criterion.

B Separating Equilibria with $\beta = 1$ and a smallest a

Suppose there exists a smallest $a > 0$, denoted by ϵ. Below we will show the existence and uniqueness of undefeated separating equilibrium in which $a_H = \epsilon > a_L = 0$ or $a_L = \epsilon > a_H = 0$ surviving the Intuitive Criterion for time consistent individuals ($\beta = 1$).

Step 1. There is a unique set of separating equilibria where $a_H = a_L + \epsilon > a_L \geq 0$ or $a_L = a_H + \epsilon > a_H \geq 0$ which survive the Intuitive Criterion.

Consider the following equilibrium: $a_H = a_L + \epsilon > a_L \geq 0$, with $\mu(a) = \begin{cases} 1 \text{ if } a = a_H \\ 0 \text{ if } a = a_L \\ p \text{ otherwise} \end{cases}$, satisfying

$$-c(\theta_L, a_L) + \int_0^{\theta_L V} (\theta_L V - k) f(k)dk > \int_0^{\bar{\theta} V} (\theta_L V - k) f(k)dk$$

and

$$-c(\theta_H, a_H) + \int_0^{\theta_H V} (\theta_H V - k) f(k)dk > \int_0^{\bar{\theta} V} (\theta_H V - k) f(k)dk.$$

So no type of individuals has incentive to deviate. Moreover, this equilibrium satisfies the Intuitive Criterion, because self-1 cannot infer the type from any off-equilibrium action.

Following the logic in Step 2 in the proof of Observation 1-a), a separating equilibrium where $a_L > a_H + \epsilon$ or $a_H > a_L + \epsilon$, if exists, does not survive the Intuitive Criterion.

Step 2. A separating equilibrium where $a_H = a_L + \epsilon > a_L > 0$ ($a_L = a_H + \epsilon > a_H > 0$) is defeated by the separating equilibrium where $a_H = a_L + \epsilon > a_L = 0$ ($a_L = a_H + \epsilon > a_H = 0$).

We show that the equilibria, denoted as σ', with $a_H = a_L + \epsilon > a_L > 0$ are defeated by the equilibrium σ with $a_H = a_L + \epsilon > a_L = 0$ in the sense of Mailath et al (1993). To see this, ϵ satisfies the following conditions: (1) No type of individuals in σ' chooses ϵ and (only) self-0 of type-H in σ chooses ϵ (satisfying D2.1 in Mailath et al, 1993); (2) self-0 of type-H is strictly better off in σ than in σ' (satisfying D2.2 in Mailath et al, 1993); (3) $\mu(\epsilon) \neq 1$ under σ', where 1 is the Bayesian belief under the strategy profile that those who choose ϵ in σ and are better off in σ than in σ' choose ϵ, those who do not choose ϵ in σ do not choose ϵ, and others may randomize (satisfying D2.3 in Mailath et al, 1993). In the equilibria σ', $\mu(\epsilon) \neq 1$ because otherwise self-0 of type-H would have incentive to deviate.

Similarly, we can show that the equilibria with $a_L = a_H + \epsilon > a_H > 0$ are defeated by the equilibrium with $a_L = a_H + \epsilon > a_H = 0$.

C Proof of Proposition 1

a) Define $a_{p,L}$ such that

$$\beta \int_0^{\beta\theta_L V} (\theta_L V - k) f(k)dk = -c(\theta_L, a_{p,L}) + \beta \int_0^{\beta\bar{\theta} V} (\theta_L V - k) f(k)dk.$$

Any $a > a_{p,L}$ cannot support a pooling equilibrium, because self-0 of type-L would have incentive to deviate to $a = 0$. For any $0 \leq a^p \leq a_{p,L}$, there is a pooling equilibrium in which both types of individuals choose a^p. To see this, let the belief system be $\mu(a) = p$ if $a \geq a^p$, and $\mu(a) = 0$ if $a < a^p$; thus, both types of individuals have no incentive to deviate.

b) For any $0 \leq a^p \leq a_{p,L}$, define $a_L^*(a^p)$ such that

$$\begin{aligned} & - c(\theta_L, a^p) + \beta \int_0^{\beta\bar{\theta}V} (\theta_L V - k) f(k) dk \\ & = -c(\theta_L, a_L^*(a^p)) + \beta \int_0^{\beta\theta_H V} (\theta_L V - k) f(k) dk \end{aligned}$$

which is equivalent to

$$\beta \int_{\beta\bar{\theta}V}^{\beta\theta_H V} (\theta_L V - k) f(k) dk = c(\theta_L, a_L^*(a^p)) - c(\theta_L, a^p).$$

$\theta_L \geq \beta\theta_H$ ensures that the left side of the last equation is positive, so $a_L^*(a^p) > a^p$. The last equation implies that

$$\begin{aligned} & \beta \int_{\beta\bar{\theta}V}^{\beta\theta_H V} (\theta_H V - k) f(k) dk \\ & > \beta \int_{\beta\bar{\theta}V}^{\beta\theta_H V} (\theta_L V - k) f(k) dk \\ & = c(\theta_L, a_L^*(a^p)) - c(\theta_L, a^p) \\ & > c(\theta_H, a_L^*(a^p)) - c(\theta_H, a^p), \end{aligned}$$

where the second inequality comes from the curvature assumptions of $c(\theta, a)$. The above is equivalent to

$$\begin{aligned} & - c(\theta_H, a^p) + \beta \int_0^{\beta\bar{\theta}V} (\theta_H V - k) f(k) dk \\ & < -c(\theta_H, a_L^*(a^p)) + \beta \int_0^{\beta\theta_H V} (\theta_H V - k) f(k) dk. \end{aligned}$$

Therefore, for any pooling equilibrium with a^p, any off-equilibrium a' which is slightly greater than $a_L^*(a^p)$ would convince self-1 that it is from type-H, because by taking a' self-0 of type-L would be worse off even if he is considered to have type-H; meanwhile, self-0 of type-H is better off by choosing a' and being considered to have type-H. The intuitive criterion rules out all the pooling equilibria.

D Proof of Proposition 2

Define $a_{s,L}$ such that

$$\beta \int_0^{\beta\theta_L V} (\theta_L V - k) f(k) dk = -c(\theta_L, a_{s,L}) + \beta \int_0^{\beta\theta_H V} (\theta_L V - k) f(k) dk$$

which is equivalent to

$$c(\theta_L, a_{s,L}) = \beta \int_{\beta\theta_L V}^{\beta\theta_H V} (\theta_L V - k) f(k) dk. \tag{2}$$

Since $\beta\theta_H V \leq \theta_L V$, the right side of the above line is positive, which ensures the existence of $a_{s,L}$.

Define $a_{s,H}$ such that

$$\beta \int_0^{\beta\theta_L V} (\theta_H V - k) f(k) dk = -c(\theta_H, a_{s,H}) + \beta \int_0^{\beta\theta_H V} (\theta_H V - k) f(k) dk$$

which is equivalent to

$$c(\theta_H, a_{s,H}) = \beta \int_{\beta\theta_L V}^{\beta\theta_H V} (\theta_H V - k) f(k) dk. \tag{3}$$

From equations (2) and (3), we have

$$c(\theta_H, a_{s,H}) - c(\theta_L, a_{s,L}) = \beta (\theta_H - \theta_L) V \int_{\beta\theta_L V}^{\beta\theta_H V} f(k) dk > 0.$$

By the monotonicity and curvature conditions for c, we thus have $a_{s,H} > a_{s,L}$.

In a separating equilibrium, the self-0 of type-L chooses zero, because otherwise by deviating to zero self-0 of type-L could be better off by reducing c and potentially enjoying a higher $\widehat{\theta}$; self-0 of type-H will not choose $a > a_{s,H}$, because otherwise self-0 of type-H has incentive to deviate to 0. For any $a^s \in [a_{s,L}, a_{s,H}]$, there is a separating equilibrium in which self-0 of type-H chooses a^s and self-0 of type-L chooses 0. (To see this, let the belief system be $\mu(a) = 1$ if $a \geq a^s$, and $\mu(a) = 0$ if $a < a^s$. With this belief system, no type has incentive to deviate under the equilibrium.) These are the unique separating equilibria.

By the Intuitive Criterion, if $a_{s,L} < a^s$, an off-equilibrium $a_{s,L} < a' < a^s$ would convince self-1 that it is from type-H, because by taking a' self-0 of type-L would be worse off even if he is considered to have type-H; meanwhile, self-0 of type-H is better off by choosing a' and being considered to have type-H. The unique separating equilibrium surviving the criterion is self-0 of type-H choosing $a_{s,L}$ and self-0 of type-L choosing 0.

E Proof of Proposition 3

Since, for all θ, $\widehat{\theta}$, and a, we have

$$U_{33}(\theta,\widehat{\theta},a) = -c_{22}(\theta,a) < 0, \qquad \text{(Strict Concavity)}$$

$$U_2(\theta,\widehat{\theta},a) = \beta\left(\theta V - \beta\widehat{\theta}V\right)\beta V f(\beta\widehat{\theta}V) > 0, \qquad \text{(Belief Monotoniciy)}$$

as $\beta < \underline{\theta}/\overline{\theta}$,

$$U_{13}(\theta,\widehat{\theta},a) = -c_{12}(\theta,a) > 0, \qquad \text{(Type Monotoniciy)}$$

$$\frac{U_3(\theta,\widehat{\theta},a)}{U_2(\theta,\widehat{\theta},a)} = -\frac{c_2(\theta,a)}{\beta\left(\theta V - \beta\widehat{\theta}V\right)\beta V f(\beta\widehat{\theta}V)} \text{ is increasing in } \theta.$$

(Single-Crossing Property)

By Mailath (1987), or more generally Mailath and von Thadden (2013), we prove the proposition.