First Draft: December, 1993
Current Draft: May, 1994
Comments appreciated.

Mental Accounts, Self-Control, and an Intrapersonal Principal-Agent Problem

David Laibson[1]

I analyze the problem of an agent with dynamically inconsistent preferences (hyperbolic discount functions) who has access to a "binding automaton:" a machine which enables the agent to perfectly commit herself to contingent rules linking observable states to observable actions. I assume that effort is not observable, generating an intra-personal principal-agent problem. In equilibrium, the agent exhibits a high marginal propensity to consume (MPC) out of effort-related income (*e.g.* labor income) and a relatively low MPC out of income which is independent of effort (*e.g.* capital gains). I interpret this as a partial explanation of mental accounts (Thaler, 1990) and of self-reward/self-punishment.

[1]This work has been supported financially by the National Science Foundation and the Alfred P. Sloan Foundation. I have benefitted from the insights of Roland Benabou, Olivier Blanchard, Matthew Rabin and participants in the MIT Macro Lunch and Theory Lunch. All mistakes should be blamed on my $t-1$ period self.

1 Introduction

When an individual's preferences are dynamically inconsistent she strictly prefers to constrain her own future choices (Strotz, 1956). Consider the following fanciful advice for a decision-maker who is looking for a way to achieve such commitment: Build a machine to follow you wherever you go, and program the machine to generate extremely aversive stimuli whenever you "misbehave." Let's call the machine a "binding automaton."

At first inspection the advice is absurd. But we do have access to social institutions which approximate the operations of the hypothesized machine. Parents, friends, supervisors, religious leaders, and spouses help us make commitments in ways which mimic the working of a binding automaton. I can promise my spouse to be home by eight. I can promise my workplace supervisor to finish a project in two weeks. Such communications implicitly generate a system of incentives which helps commit me to honor my promise.

In general, any organization/group in which I am a member has a seemingly endless and often unspoken list of expectations/rules/norms to which I commit when I join the group. For example, my friends see me wearing a very luxurious new coat, and ask me about it. If I acknowledge that I spent $400 dollars on the coat they call me a spendthrift and effectively censure me. If I tell them that I found it on sale, they complement me on my good taste. Hence, membership in my social group is associated with strong incentives to not purchase luxurious clothing at retail prices.

This analysis suggests a theory of organizations and groups based on the commitment value of these institutions. Examples would include firms and schools (committing me to be productive), religions and marriage (committing me to be "virtuous"), twelve-step groups (committing me to be abstemious), etc. With these examples in mind, it may be reasonable to assume that individuals have access to social institutions which act effectively as binding automata. This is a strong assumption, but one which is worth

considering as an important benchmark. The analysis which follows takes this assumption as a starting point and sees where it leads. Specifically, I analyze the behavior of an agent who has access to a binding automaton which enables her to perfectly commit to any contingent rule linking observable states to observable actions.

In a world where *all* states and actions are observable, the binding automaton assumption dramatically simplifies analysis. Actions are simply the optimal contingent rules from the perspective of the self that builds the binding automaton. However, in an economy where some actions and/or states are not observable the analysis is more complex, and that is the world which I will analyze. Specifically, I assume that the binding automaton can not directly observe effort.

The problem that I consider is an intra-personal principal-agent problem. The principal is the early self who builds the binding automaton. The agent is the later self who exerts effort in response to the incentive system created by the binding automaton. I assume that both the principal and the agent have a hyperbolic discount function.[2] This discount structure implies that the principal prefers relatively less consumption and relatively more effort during the agent's period of control than does the agent.

This intra-personal principal-agent problem can be compared and contrasted with the standard principal-agent problem from the industrial organization literature (*e.g.*, Shavell (1979), Holmström (1979), Grossman and Hart (1983)). In both problems effort is not observable, and the principal sets up an incentive system to motivate effort. However, in the intra-personal principal-agent model the utility of the principal and the agent are linked in a way which bears no resemblance to the utility relationship in the standard principal-agent model.

In the intra-personal principal-agent problem, the principal is forced to

[2]See my first two chapters for an overview of hyperbolic discount functions.

strike a balance between consumption smoothing, which the principal wants, and instantaneous gratification, which is used to motivate the agent to exert high effort. This fundamental tension explains many heretofore puzzling anomalies, including self-reward/self-punishment and some mental accounts. The principal, or early self, would like the later self to exert high effort *and* smooth consumption. But in order to extract high effort, the principal has to give the later self a strong incentive. The incentive takes the form of the following norm: if you (the later self) obtain a good-effort related outcome, then you can splurge. Hence, the equilibrium path is characterized by a high marginal propensity to consume out of labor income. This is self-reward. However, the early self does not need to reward the later self for good outcomes that are independent of effort. So when asset stocks are high, consumption responds modestly (according to the wishes of the early self). Hence on the equilibrium path, the marginal propensity to consume out of assets is low when compared to the marginal propensity to consume out of labor income. This pattern replicates some of the mental accounting behavior documented by Richard Thaler (1990).

The body of the paper formalizes these claims. Section 2 lays out the model. Equilibrium outcomes are characterized in Section 3. A numerical example is illustrated in Section 4. Section 5 concludes with a critical evaluation and a discussion of ongoing work.

2 An Intra-personal Principal-Agent Model

The model has three critical components. First, the individual is assumed to have a hyperbolic discount function. This sets up an intra-personal conflict. Second, the individual has access to a costless binding automaton. This implies that the initial self can perfectly commit the observable actions of future selves. These commitments take the form of contingent rules, where the contingencies are based on observable states. Third, I assume that effort

is not observable, but effort is positively correlated with good outcomes. The details of the model follow.

The individual makes decisions over three periods, $t \in \{1, 2, 3\}$. I adopt the semantic convention used in my previous chapters, and refer to self t as the self in control at time t. During period 1 the binding automaton is built by self 1 to maximize the welfare of self 1. In period 2, self 2 chooses effort, receives labor income (a stochastic function of effort), and consumes according to the instructions of the binding automaton. In period 3, self 3 consumes whatever assets are left.[3]

The individual is assumed to have a hyperbolic discount function. I capture the properties of the hyperbolic discount function by assuming that the one period discount factor is $\beta\delta$, $(0 < \beta < 1)$, and assuming that the two-period discount factor is $\beta\delta^2$. This implies that the discount rate is falling. The utility functions of the three selves are given by,

$$\begin{array}{lrcrcr} \text{Self 1:} & u_1 & + & \beta\delta\left[u(c_2) - e\right] & + & \beta\delta^2 u(c_3) \\ \text{Self 2:} & & & \left[u(c_2) - e\right] & + & \beta\delta u(c_3) \\ \text{Self 3:} & & & & & u(c_3) \end{array}$$

where u_1 is a constant, $u' > 0$, $u'' < 0$. The production technologies in the economy are very simple. Effort (chosen in period 2) takes one of two values, $e \in \{\underline{e}, \overline{e}\}$, $\underline{e} < \overline{e}$. Income (realized in period 1) is also drawn from a two-point set: $y \in \{y_L, y_H\}$, $y_L < y_H$. The distribution of income is a function of the chosen effort level. Specifically,

$$Prob(y_H|\overline{e}) = \overline{p} > \underline{p} = Prob(y_H|\underline{e}),$$

[3]It is possible to also let self 3 consume according to rules of the binding automaton, but those rules will generally dictate that self 3 consume all remaining assets anyway. The only exception to this pattern arises when the optimal contingent rules in period 2 constitute a border solution. In that case it may be *ex ante* optimal to instruct self 3 to consume less than the remaining asset stock contingent on a low income realization in period 2.

which implies that $Prob(y_L|\overline{e}) = 1 - \overline{p}$, and $Prob(y_L|\underline{e}) = 1 - \underline{p}$. Finally, the gross interest rate, R characterizes the storage technology. For simplicity, I set $R = 1$.

We are now in a position to discuss explicitly the construction of the binding automaton. The only observable state in period 2 will be the realized income level, y. So the consumption rule for period 2 can only be a function of y. Since y can only take on two values, it is possible to represent the consumption rule as an ordered pair: $\{c_L, c_H\}$, where c_L is the consumption level when y_L is realized, and c_H is the consumption level when y_H is realized.

3 Equilibrium

Self 2's problem must be solved first. Let, $W(c_L, c_H) \equiv$

$$(\overline{p} - \underline{p})\left[u(c_H) + \beta\delta u(y_H - c_H) - u(c_L) - \beta\delta u(y_L - c_L)\right] - (\overline{e} - \underline{e}),$$

which is the difference between self 2's payoff given $e = \overline{e}$ and self 2's payoff given $e = \underline{e}$. Self 2 must select an effort level e^*. She sets $e^* = \overline{e}$ if $W > 0$, $e^* = \underline{e}$ if $W < 0$, and is assumed to choose whatever effort level is preferred by self 1 if $W = 0$.

Self 1's decision problem is less simple. Let $U(c_L, c_H|\overline{e}) \equiv$

$$\overline{p}\left[u(c_H) + \beta\delta u(y_H - c_H)\right] + (1 - \overline{p})\left[u(c_L) + \beta\delta u(y_L - c_L)\right] - \overline{e},$$

and let $U(c_L, c_H|\underline{e})$ represent the same object with $\overline{p}$ replaced by $\underline{p}$ and $\overline{e}$ replaced by $\underline{e}$. Hence, $U(c_L, c_H|\overline{e})$ is the payoff to self 1, given $e = \overline{e}$, and $U(c_L, c_H|\underline{e})$ is the payoff to self 1, given $e = \underline{e}$.

Self 1's decision process requires self 1 to solve two subproblems and compare the solutions.

$$\text{I.} \quad \max_{\{c_L, c_H\}} \; U(c_L, c_H|\overline{e}) \quad \text{s.t.} \; W(c_L, c_H) \geq 0$$

$$\text{II.} \quad \max_{\{c_L, c_H\}} \; U(c_L, c_H|\underline{e}) \quad \text{s.t.} \; W(c_L, c_H) \leq 0$$

Let C^I (C^{II}) represent the set of ordered pairs which are the solutions to program I (II). Let C^* represent the set of ordered pairs which are the solutions to self 1's global problem.

$$C^* = \begin{cases} C^I & \text{if } U(C^I|\overline{e}) \geq U(C^{II}|\underline{e}) \\ C^{II} & \text{otherwise} \end{cases}$$

3.1 Self 1's first-best solution

The first goal of this section is to simplify self 1's decision problem. To do this it is helpful to discuss a benchmark case: the first-best solution from the perspective of self 1. Consider the problem in which self 1 chooses the ordered pair $\{c_L, c_H\}$ and also directly chooses the effort level, e. It is trivial to show that there exists a unique ordered pair, $\hat{C} = \{\hat{c}_L, \hat{c}_H\}$ which solves this first-best problem. This ordered pair is determined by the first-order conditions:

$$u'(\hat{c}_H) = \delta u'(y_H - \hat{c}_H),$$

$$u'(\hat{c}_L) = \delta u'(y_L - \hat{c}_L).$$

The effort level, $\hat{e}$, which solves self 1's first-best problem is given by,

$$\hat{e} = \begin{cases} \overline{e} & \text{if } U(\hat{C}|\overline{e}) \geq U(\hat{C}|\underline{e}) \\ \underline{e} & \text{otherwise} \end{cases}$$

The following lemma will be used in several of the results which follow.

Lemma 1: $y_H - \hat{c}_H > y_L - \hat{c}_L$.

Proof: Suppose $y_H - \hat{c}_H \leq y_L - \hat{c}_L$, and look for a contradiction. By FOC's of first-best problem,

$$u'(\hat{c}_H) = \delta u'(y_H - \hat{c}_H) \geq \delta u'(y_L - \hat{c}_L) = u'(\hat{c}_L),$$

where the inequality follows from the assumption, $y_H - \hat{c}_H \leq y_L - \hat{c}_L$. The inequality $u'(\hat{c}_H) \geq u'(\hat{c}_L)$ implies $\hat{c}_H \leq \hat{c}_L$. So,

$$\hat{c}_H + (y_H - \hat{c}_H) \leq \hat{c}_L + (y_L - \hat{c}_L),$$

implying, $y_H \leq y_L$, which is a contradiction. □

3.2 Simplifying self 1's problem

We are now ready to simplify self 1's decision rule. In particular, it is possible to solve self 1's problem without solving program II. The following lemma is used to prove this result.

Lemma 2: *If* $U(C^I|\bar{e}) < U(C^{II}|\underline{e})$, *then* $C^* = \hat{C}$.

Proof: Suppose $U(C^I|\bar{e}) < U(C^{II}|\underline{e})$, and $C^* \neq \hat{C}$, and look for contradiction. Recall that in general C^* is a solution *set*. It is easy to show that $C^* \neq \hat{C}$ implies that $C^* \not\ni \hat{C}$. Continuing with the proof, note that $W(\hat{C})$ must be greater than zero, (since $C^* = C^{II}$, C^{II} is the solution set of program II, $C^* \not\ni \hat{C}$, and $U(\hat{C}|\underline{e}) \geq U(C^*|\underline{e})$). $W(\hat{C}) > 0$ implies that,

$$U(\hat{C}|\bar{e}) \leq U(C^I|\bar{e}) \leq U(C^*|\underline{e}) \leq U(\hat{C}|\underline{e}).$$

Combining, $W(\hat{C}) > 0$ with $U(\hat{C}|\overline{e}) \leq U(\hat{C}|\underline{e})$ yields,

$$u(y_L - c_L) \geq u(y_H - c_H),$$

which contradicts lemma 1. □

The following proposition shows that self 1's optimal decision can be calculated without solving program II.

Proposition 1: *Let*

$$C^{**} = \begin{cases} C^I & \text{if } U(C^I|\overline{e}) \geq U(\hat{C}|\underline{e}) \\ \hat{C} & \text{otherwise} \end{cases}$$

Then $C^{**} = C^*$.

Proof: First, suppose $U(C^I|\overline{e}) < U(C^{II}|\underline{e})$. Then $C^* = C^{II}$, and by Lemma 2, $C^{II} = \hat{C}$. Hence, for this case the proposition is true. Now WLOG, assume, $U(C^I|\overline{e}) \geq U(C^{II}|\underline{e})$. If $C^{II} = \hat{C}$, the proposition is true. So WLOG, assume, $C^{II} \neq \hat{C}$. This implies, $W(\hat{C}) > 0$, which implies that $C^I = \hat{C}$. Hence,

$$U(C^I|\overline{e}) = U(\hat{C}|\overline{e}) > U(\hat{C}|\underline{e}),$$

where the last inequality is derived by combining Lemma 1 with $W(\hat{C}) > 0$. Hence, $C^{**} = C^I$, completing the proof. □

3.3 Characterizing the solution

This subsection characterizes the set of solutions to self 1's problem. This characterization is useful for two reasons. First, it helps to develop intuition about the solutions. Second, it enables me to prove that second order conditions are satisfied. Program 1 is not concave over the entire solution space.

However, the solutions of the problem can be shown to exist in a subspace in which sufficient conditions are satisfied. I will return to these issues in the next subsection. I use the following notation in the claims below:

$$U_H \equiv \frac{\partial U}{\partial c_H}, \quad U_L \equiv \frac{\partial U}{\partial c_L}, \quad W_H \equiv \frac{\partial W}{\partial c_H}, \quad W_L \equiv \frac{\partial W}{\partial c_L}.$$

Proposition 2: *In equilibrium* $u'(c_H) \geq \beta\delta u'(y_H - c_H)$.

Proof: If $C^* = \hat{C}$, the proposition follows immediately, (since $\beta < 1$). So, WLOG, assume $C^* \neq \hat{C}$. This implies (by Proposition 1) that $C^* = C^I$. The Kuhn-Tucker necessary conditions associated with program I are given below:

$$U_H + \lambda W_H = 0,$$

$$U_L + \lambda W_L = 0,$$

$$\lambda \geq 0, \quad W \geq 0, \quad \lambda W = 0.$$

Note that the proposition is equivalent to the claim that in equilibrium $W_H \geq 0$. Assume that in equilibrium $W_H < 0$, and look for a contradiction. Note that $U_H = W_H - (1-\beta)\delta u'(y_H - c_H)$, so $W_H < 0$ implies $U_H < 0$, which implies $\lambda < 0$, which contradicts the necessary conditions. □

Proposition 3: *In equilibrium* $u'(c_H) \leq \delta u'(y_H - c_H)$.

Proof: Using the same argument as above, assume WLOG $C^* \neq \hat{C}$, and hence, $C^* = C^I$. Recall the Kuhn-Tucker conditions associated with program I. Note that $W_H \geq 0$ (by Proposition 2), and $\lambda \geq 0$ implies $U_H \leq 0$, which completes the proof. □

Lemma 3: *In equilibrium,* $u'(c_L) \geq \delta u'(y_L - c_L)$ *or* $u'(c_L) < \beta\delta u'(y_L - c_L)$.

Proof: Using the same argument as above, assume WLOG $C^* \neq \hat{C}$, and hence, $C^* = C^I$. Recall the Kuhn-Tucker conditions associated with program I. Note that $u'(c_L) < \delta u'(y_L - c_L)$ implies $U_L < 0$, which implies, $W_L > 0$, (since $\lambda \geq 0$). □

Lemma 4: *In equilibrium,* $u'(c_L) \geq \beta\delta u'(y_L - c_L)$.

Proof: Using the same argument as above, assume WLOG $C^* \neq \hat{C}$, and hence, $C^* = C^I$. Fix any equilibrium c_L. Suppose, $u'(c_L) < \beta\delta u'(y_L - c_L)$, and look for a contradiction. Note that $u(x) + \beta\delta u(y_L - x)$ is concave in x with a maximum at $x : u'(x) = \beta\delta u'(y_L - x)$. Note that $u(0) + \beta\delta u(y_L) \leq u(y_L) + \beta\delta u(0)$. So there exists a $\tilde{c}_L < c_L$ s.t.

$$u(\tilde{c}_L) + \beta\delta u(y_L - \tilde{c}_L) = u(c_L) + \beta\delta u(y_L - c_L).$$

Note that $u(y_L - \tilde{c}_L) > u(y_L - c_L)$. Combining this observation with the previous line yields,

$$u(\tilde{c}_L) + \delta u(y_L - \tilde{c}_L) > u(c_L) + \beta\delta u(y_L - c_L),$$

which implies that self 1 is made strictly better off by switching from c_L to $\tilde{c}_L$, which violates the original equilibrium assumption. □

Proposition 4: *In equilibrium* $u'(c_L) \geq \delta u'(y_H - c_L)$.

Proof: The proposition follows from Lemma 3 and Lemma 4. □

Proposition 5: *In equilibrium,*

$$\frac{c_H - c_L}{y_H - y_L} \leq 1.$$

Proof: If $C^* = \hat{C}$ then the Proposition follows from Lemma 1. WLOG assume $C^* = C^I \neq \hat{C}$. This implies that $W(C^*) = 0$. Note that any perturbation of C^* must make self 1 no better off. Consider perturbations to C^* which lead self 2 to choose $\underline{e}$ instead of $\bar{e}$. Such perturbations are possible since $W(C^*) = 0$. Optimality of C^* requires that,

$$U(C^*|\bar{e}) - U(C^*|\underline{e}) \geq 0.$$

Subtracting $W(C^*) = 0$ from the LHS of this expression, yields,

$$\delta(1-\beta)(\bar{p} - \underline{p})\left[u(y_H - c_H) - u(y_L - c_L)\right] \geq 0.$$

This implies that $(y_H - c_H) \geq (y_L - c_L)$ which completes the proof. □

3.4 Sufficient Conditions

It is now possible to derive a sufficiency theorem.

Proposition 6: *There exists at most one solution to program I. If a solution exists, it is in the region described by Propositions*

2-4, and it is the only point in this region which satisfies the Kuhn-Tucker conditions of program I.

Proof: The proposition is trivial to confirm if $W(\hat{C}) \geq 0$. WLOG assume $W(\hat{C}) < 0$. Then at any solution, $W = 0$. So the solution set of program I (under these assumptions) is equivalent to the solution set of the following program, (in which the constraint binds).

$$\begin{aligned} \text{IB.} \quad & \max_{\{c_L, c_H\}} \quad U(c_L, c_H|\bar{e}) \\ & \text{s.t.} \quad W(c_L, c_H) = 0 \end{aligned}$$

Define a subspace S of the non-negative orthant of $\Re^3$, such that elements of S are ordered triplets, $\{c_L, c_H, \lambda\}$ which satisfy the properties,

$$\delta u'(y_L - c_L) \leq u'(c_L),$$

$$\beta\delta u'(y_H - c_H) \leq u'(c_H) \leq \delta u'(y_H - c_H),$$

$$\lambda \leq \frac{1-\bar{p}}{\bar{p}-\underline{p}}.$$

Note that by Propositions 2-4 and the Kuhn-Tucker conditions of program I, any solution to program I must be an element of S. Moreover, since program I and program IB have the same solution set and the same Kuhn-Tucker conditions, all solutions of program IB must also be in S.

The next step of the proof is to show that the bordered Hessian associated with program IB has a positive determinant in S. Let,

$$U_{HH} \equiv \frac{\partial^2 U}{\partial c_H^2}, \quad U_{LL} \equiv \frac{\partial^2 U}{\partial c_L^2}, \quad U_{HL} \equiv \frac{\partial^2 U}{\partial c_H \partial c_L}.$$

and represent the second derivatives of W in an analogous way. Let λ be the Lagrange multiplier associated with the constraint in program IB. Then

the bordered Hessian of program IB is,

$$H \equiv \begin{bmatrix} 0 & W_H & W_L \\ W_H & U_{HH} + \lambda W_{HH} & U_{HL} + \lambda W_{HL} \\ W_L & U_{HL} + \lambda W_{HL} & U_{LL} + \lambda W_{LL} \end{bmatrix}$$

Note that $U_{LH} = 0$, and $W_{HL} = 0$. So the determinant of the bordered Hessian is,

$$|H| = -\left[W_H^2(U_{LL} + \lambda W_{LL}) + W_L^2(U_{HH} + \lambda W_{HH})\right].$$

Note that $(U_{HH} + \lambda W_{HH}) < 0$ in S, (since $U_{HH} < 0$, $\lambda \geq 0$, and $W_{HH} < 0$ in S). Hence, to show $|H| > 0$, it is sufficient to show $U_{LL} + \lambda W_{LL} =$

$$(1 - \overline{p})\left[u''(c_L) + \delta u''(y_L - c_L)\right] - \lambda(\overline{p} - \underline{p})\left[u''(c_L) + \beta\delta u''(y_L - c_L)\right] < 0,$$

in S. This inequality follows from the properties, $\beta < 1$, and $\lambda \leq \frac{1-\overline{p}}{\overline{p}-\underline{p}}$.

Hence, the determinant of the bordered Hessian is positive in S, so there exists a unique point in S which satisfies the Kuhn-Tucker conditions of program IB. (Note, there must exist at least one point in S which satisfies the Kuhn-Tucker conditions, since S contains all solutions of program IB.) Since all solutions of program IB satisfy the Kuhn-Tucker conditions, and a unique point in S satisfies the Kuhn-Tucker conditions, and all solutions of program IB are in S, there exists a unique solution of program IB. Hence, program I must also have a unique solution.

Note that the solution to program IB satisfies the Kuhn-Tucker conditions of program I. Now it only remains to show that the Kuhn-Tucker conditions of program I admit no other solutions in S. Let C be the unique maximum of the two programs. Let C' be any point in S which satisfies the Kuhn-Tucker conditions of program I. If $W(C') = 0$ then C' also satisfies

the Kuhn-Tucker conditions of program IB, implying that $C' = C$ (since I have shown that C is the only point in S which satisfies the Kuhn-Tucker conditions of program IB). So WLOG assume that $W(C') > 0$. Then $\lambda = 0$, and $C' = \hat{C}$, contradicting the assumption $W(\hat{C}) < 0$. □

3.5 Comparative Statics

This model suggests a natural measure of the marginal propensity to consume out of labor income:

$$MPC^y = \frac{c_H - c_L}{y_H - y_L}.$$

Note that β measures the degree of congruence between the interests of self 1 and self 2. When $\beta = 1$ those interests are perfectly aligned, and the principal-agent problem collapses to the first-best problem of self 1. As β goes to zero the interests of self 1 and self 2 are maximally unaligned. At the limit, self 2 desires complete instantaneous gratification.

Proposition 7: *Let $C(\beta)$ be the (unique) solution to self 1's problem, given a particular value of β. If $C(\beta) \neq \hat{C}$,*

$$\frac{\partial MPC^y(\beta)}{\partial \beta} < 0.$$

Proof: $C(\beta) \neq \hat{C}$ implies that $W(C(\beta)) = 0$, so the inequality constraint in program I binds. Hence, in equilibrium the following equations hold (derived from the Kuhn-Tucker conditions): $U_H W_L = U_L W_H$, and $W = 0$. Applying the implicit differentiation theorem and eliminating zero terms yields,

$$\frac{\partial c_H}{\partial \beta} = \frac{W_L[U_H W_{L\beta} - U_L W_{H\beta}] - W_\beta[U_H W_{LL} - U_{LL} W_H]}{|H|},$$

$$\frac{\partial c_L}{\partial \beta} = \frac{-W_H[U_H W_{L\beta} - U_L W_{H\beta}] + W_\beta[U_{HH} W_L - U_L W_{HH}]}{|H|}.$$

Recall (from previous proof) that $|H|$ is the determinant of the bordered Hessian. Combining the previous two equations yields, $\frac{\partial(c_H - c_L)}{\partial \beta} =$

$$\frac{(W_L + W_H)[U_H W_{L\beta} - U_L W_{H\beta}] - W_\beta[U_H W_{LL} - U_{LL} W_H + U_{HH} W_L - U_L W_{HH}]}{|H|}$$

I've already shown (see previous proof) that $|H|$ is positive at all optima. Hence, it is sufficient to sign the numerator of this expression. I proceed term by term.

Note that,

$$\begin{aligned}
\frac{-W_L}{W_H} &\geq \frac{-W_L}{W_H} \cdot \frac{u'(y_H - c_H)}{u'(y_L - c_L)} \\
&= \frac{\frac{u'(c_L) - \beta\delta u'(y_L - c_L)}{u'(y_L - c_L)}}{\frac{u'(c_H) - \beta\delta u'(y_H - c_H)}{u'(y_H - c_H)}} \\
&= \frac{\frac{u'(c_L)}{u'(y_L - c_L)} - \beta\delta}{\frac{u'(c_H)}{u'(y_H - c_H)} - \beta\delta} \\
&> 1
\end{aligned}$$

where the first inequality follows from Proposition 5, and the last inequality follows from Proposition 3 and Proposition 4. Note that $W_H > 0$ by Proposition 2. Multiplying the last line by W_H yields,

$$W_H + W_L < 0.$$

Note that,

$$\begin{aligned}\frac{-U_L W_{H\beta}}{U_H W_{L\beta}} &= \frac{-W_L W_{H\beta}}{W_H W_{L\beta}} \\ &= \frac{-\frac{u'(c_L)-\beta\delta u'(y_L-c_L)}{u'(y_L-c_L)}}{\frac{u'(c_H)-\beta\delta u'(y_H-c_H)}{u'(y_H-c_H)}} \\ &< -1\end{aligned}$$

The first inequality follows from substitution of the Kuhn-Tucker conditions, and the last inequality follows from the arguments made in the previous derivation. Note that $U_H W_{L\beta} < 0$ by Proposition 3. Multiplying the last line by $U_H W_{L\beta}$ yields,

$$U_H W_{L\beta} - U_L W_{H\beta} > 0.$$

Note that,

$$\begin{aligned}\frac{U_H W_{LL}}{U_{LL} W_H} &= \frac{-U_L W_{LL}}{W_L U_{LL}} \\ &= \frac{u'(c_L) - \delta u'(y_L - c_L)}{u'(c_L) - \beta\delta u'(y_L - c_L)} \cdot \frac{u''(c_L) + \beta\delta u''(y_L - c_L)}{u''(c_L) + \delta u''(y_L - c_L)} \\ &> 1\end{aligned}$$

The first inequality follows from substitution of the Kuhn-Tucker conditions, and the last inequality follows from $\beta < 1$, and Proposition 4. Note that

$U_{LL}W_H < 0$, by Proposition 2. Multiplying the last line by $U_{LL}W_H$ yields,

$$U_H W_{LL} - U_{LL} W_H > 0.$$

The remaining terms can be signed directly. Note that $W_\beta > 0$ by Proposition 5, $U_L > 0$ by Proposition 4, $W_L < 0$ by Proposition 4, and $U_{HH}, W_{HH} < 0$. So,

$$U_{HH} W_L - U_L W_{HH} > 0.$$

Together these observations imply that $MPC^y(\beta)$ is falling in β. □

This long proof establishes a very simple result. As self 2's interests move closer to self 1's interests, self 1 needs to sanction less instantaneous gratification to motivate self 2. Put differently, self-reward is used less and less for self-motivation as β goes to unity.

3.6 Mental Accounts

I am now in a position to return to the discussion of mental accounts. The goal of this subsection is to propose a meaningful way of comparing the MPC out of labor income—MPC^y—with the MPC out of assets, henceforth represented as MPC^A. The first order of business is to define MPC^A.

To simplify analysis I will propose a definition of MPC^A which is independent of the effort incentive problem. Modify the earlier intra-personal principal-agent problem by setting $\bar{p} = \underline{p}$. This modified problem is a limiting case of the original problem. Note that the modified problem has no meaningful effort decision. Income in the modified problem is uncorrelated with the effort decision. So the agent trivially chooses low effort. In this setting income can be interpreted as asset windfalls (*e.g.* capital gains). Define

MPC^A to be the MPC which arises in the modified problem. Specifically,

$$MPC^A = \frac{c_H - c_L}{y_H - y_L},$$

where c_H and c_L are the equilibrium contingent consumption levels associated with the modified problem.[4]

Proposition 8: *Let MPC^A be the equilibrium MPC of the limiting principal-agent problem with $\overline{p} = \underline{p}$. Let MPC^v_{FB} be the MPC associated with the first-best solution of the original principal-agent problem. Let $MPC^v(1)$ be the equilibrium MPC of the original principal-agent problem with $\beta = 1$. Let $MPC^v(\beta)$ be the equilibrium MPC of the original principal-agent problem with $\beta < 1$. Then, if C^* is the equilibrium in the original principal-agent problem, and $C^* \neq \hat{C}$,*

$$MPC^A = MPC^v_{FB} = MPC^v(1) < MPC^v(\beta).$$

Proof: It is straightforward to confirm that the agency problems associated with MPC^A, MPC^v_{FB}, and $MPC^v(1)$, all have solution $\{c_H, c_L\} = \hat{C}$. The first two inequalities follow immediately from this observation. The last inequality is implied by Propositions 3 and 4, and $C^* \neq \hat{C}$. □.

[4]Other sensible definitions are possible. In particular, I originally worked with the definition

$$MPC^A = \omega \frac{\partial c_H}{\partial \Delta} + (1-\omega)\frac{\partial c_L}{\partial \Delta},$$

where ω is a weighting function (*e.g.* $\omega = \overline{p}$, or $\omega = \frac{1}{2}$), and Δ represents an income component which is in both y_H and y_L. This approach generated the same qualitative results as the one pursued in the paper, but with far less clarity and simplicity.

4 An illustration

I have shown that the MPC out of labor income (*i.e.* income which is positively correlated with effort) is higher than the MPC out of asset income (*i.e.* income which is uncorrelated with effort). But I haven't discussed the magnitude of this difference. The following example provides an arbitrary illustration of the magnitude of these effects. The example is calibrated by setting $u(\cdot) = \ln(\cdot)$, $\delta = 1$, $y_H = 6$, $y_L = 5$, $\overline{p} = \frac{2}{3}$, $\underline{p} = \frac{1}{3}$, and $\overline{e} - \underline{e} = .116$. All of these parameters and specifications were chosen independently, except $\overline{e} - \underline{e}$, which was chosen so that there would be a range of β values over which self 1 would induce self 2 to select the high effort level.

Figure 1 graphs c_H and c_L as functions of β. The unit interval of β values can be broken down into three subintervals of interest: $\beta \in [0, .1002]$, $\beta \in [.1002, .9087]$, and $\beta \in [.9087, 1]$. I will refer to these as intervals A, B, and C. In interval A, β is close to zero, and the interests of self 1 and self 2 are highly divergent. Self 1 would like self 2 to exert high effort, but setting up an incentive scheme to induce self 2 to do so is too costly from the perspective of self 1. So self 1 choose $C^* = \hat{C}$ and self 2 chooses $e^* = \underline{e}$. In interval B, β takes on intermediate values. Now, from self 1's perspective it is desirable to set up an incentive scheme to induce self 2 to exert high effort. Because self 2's interests are still sufficiently divergent from self 1's interest, self 2 must be rewarded to motivate a high effort choice. Hence in interval B, $c_H^* > \hat{c}_H$, and $c_L^* < \hat{c}_L$. In interval C, β is sufficiently close to one that self 1 does not need to create any (extra) incentive to motivate self 2 to choose $\overline{e}$. So in this region, $C^* = \hat{C}$, and $e^* = \overline{e}$.

Figure 2 graphs, $MPC^y(\beta) = \frac{c_H(\beta) - c_L(\beta)}{y_H - y_L}$. MPC^y should be contrasted with MPC^A; the latter is equal to .5 for all β values. Finally, note that the drop in MPC^y at low β values is a consequence of the discrete effort assumption. I conjecture that a model with continuous effort choice will generate an equilibrium MPC^y which rises everywhere as β falls. Preliminary work

Figure 1: cH and cL as a function of Beta

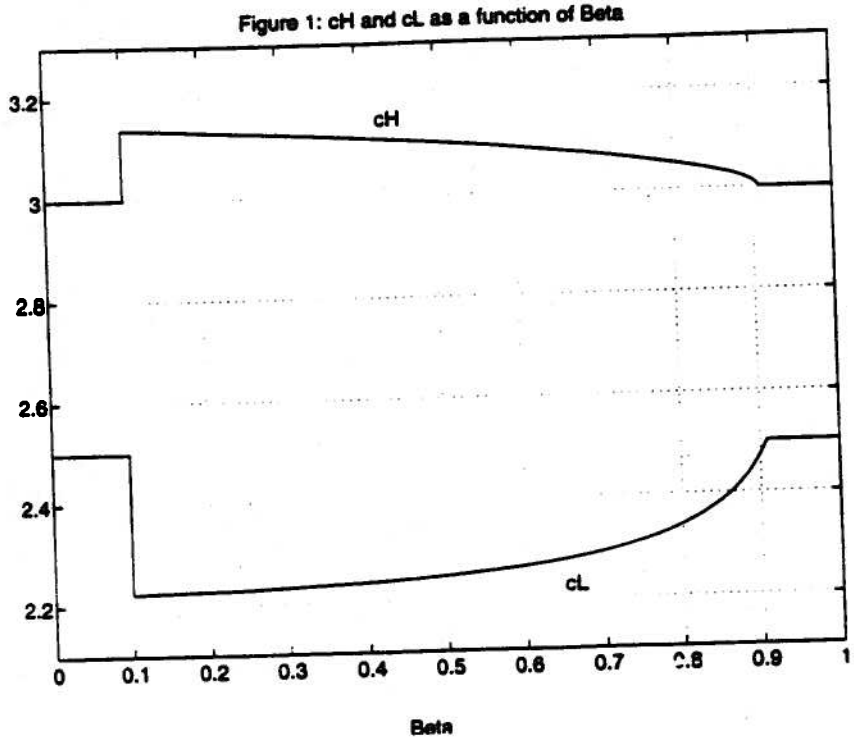
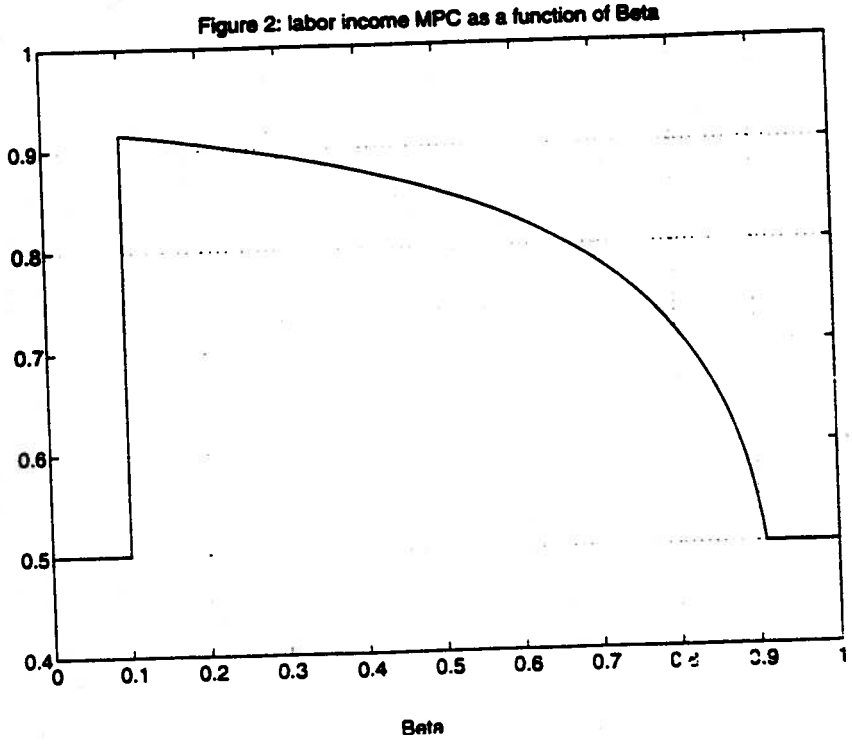


Figure 2: labor income MPC as a function of Beta



with such a model supports this observation.

5 Evaluation

The intrapersonal principal agent problem in these notes explains mental accounts in a novel way. In my story, mental accounts represent a sophisticated tradeoff between the desire for consumption smoothing and the need to motivate effort with instantaneous gratification/punishment. I acknowledge that not all mental account phenomena fit naturally into this paradigm. In particular, the most problematic mental account phenomenon for my intra-personal agency model is the high measured MPC out of exogenous, liquid wealth windfalls (see Thaler, 1990). Such behavior can be shoe-horned into my intra-personal agency model in either of two ways. First, one can assume that no liquid wealth shocks are truly independent of effort, or at least that if such exogenous shocks do exist, they are sufficiently rare or difficult to identify that we do not even bother creating norms to handle them. Second, one can assume that the high measured MPC out of liquid wealth shocks reflects an incentive scheme that was historically highly successful but has ceased to be useful due to a weakening of the connection between recent effort and liquid wealth. There may have been a time when all liquid wealth shocks were effort-related—*e.g.*, in a hunter-gatherer society—and that is when these norms evolved.

While I find both of these stories intriguing, I believe that my intra-personal agency model does not satisfactorily explain the high measured MPC out of exogenous liquid wealth windfalls. However, any model with liquidity constraints (endogenous or exogenous), can explain the liquid wealth anomaly (*e.g.* see the second chapter of this thesis). Hence, I am comfortable separating the liquid wealth anomaly from the body of other mental accounting phenomena that my intra-personal agency model does readily explain: a high MPC out of income/wealth that is effort related (*e.g.* labor

income) and a relatively low MPC out of income/wealth that is independent of effort (*e.g.* capital gains). For example, my agency model explains why passive investors have a lower MPC out of capital gains than active investors. The agency model explains why students reward themselves when they get back a successful exam. The agency model explains why shoppers reward themselves for finding a needed item on sale (by splurging the "saved" money on something frivolous). The agency model explains why many people punish themselves when something "bad" happens, like losing an airplane ticket or getting fined for speeding.

Future work on my intra-personal agency model of mental accounts will focus in two areas. First, I hope to extend the model to a multi-period framework. Such an extension will make the model more closely map observed mental account phenomena by increasing the gap between the MPC out of effort-independent wealth and the MPC out of effort-related wealth. Second, I hope to weaken my assumption about the effectiveness of the binding automaton. I have assumed that the binding automaton can be used to perfectly commit the agent to *any* contingent rule linking observable actions to observable states. Weakening this assumption is analogous to weakening the complete contracts assumption in the standard principal-agent literature. Preliminary work suggests that such a weakening will pave the way for a rich theory of organizations and norms based on the value of commitment.

References

Ainslie, George W. (1992) *Picoeconomics,* Cambridge: Cambridge University Press.

Grossman, Sanford, and Oliver Hart (1983) "An Analysis of the Principal-Agent Problem." *Econometrica,* 51, 7-45.

Holmström, B. (1979) "Moral Hazard and Observability." *Bell Journal of Economics,* 10, 74-91.

Laibson, David I. (1992) "Self-control and Saving." MIT mimeo.

Laibson, David I. (1993) "Golden Eggs and Hyperbolic Discounting." MIT mimeo.

Shavell, Steve (1979) "Risk Sharing and Incentives in the Principal and Agent Relationship." *Bell Journal of Economics,* 10, 55-73.

Strotz, Robert H. (1956) "Myopia and Inconsistency in Dynamic Utility Maximization." *Review of Economic Studies,* 23, 165-180.

Thaler, Richard H. (1990) "Saving, Fungibility, and Mental Accounts." *Journal of Economic Perspectives,* 4:1, 193-205.