The Brain as a Hierarchical Organization *

Isabelle Brocas
USC and CEPR

Juan D. Carrillo
USC and CEPR

This version: September 2006

Abstract

Inspired by recent neuroscience evidence, we model the brain as a dual-system organization subject to three conflicts: asymmetric information, temporal evaluation and incentive salience. We show that, under the first two conflicts, the uninformed system proposes a self-disciplining rule of the type "work more today to consume more today". Under the first and third conflict, the optimal rule becomes a simple, non-intrusive precept of the type "consume what you want but don't abuse". We discuss the behavioral implications of these rules for choice bracketing and expense tracking, consumption over the life-cycle and the rate of time-preference.

*We thank Roland Bénabou, Colin Camerer, Stefano DellaVigna, Christian Hellwig, Botond Koszegi, George Loewenstein, John O'Doherty, Ignacio Palacios-Huerta, Drazen Prelec, Matthew Rabin, Antonio Rangel, John Riley, Hersh Shefrin, Bill Zame and seminar participants at USC, Princeton, Columbia, Toulouse, Stanford SITE, Caltech, UC Davis, SWET (Arizona State University), UC Los Angeles and UC Berkeley for comments and suggestions. We are especially thankful to Antoine Bechara and Paul Glimcher for their guidance and patience. Address for correspondence: Isabelle Brocas or Juan D. Carrillo, Department of Economics, University of Southern California, 3620 S. Vermont Ave., Los Angeles, CA - 90089-0253, e-mail: <brocas@usc.edu> or <juandc@usc.edu>.

"The heart has its reasons which reason knows nothing of"
(Blaise Pascal (1670), *Les Pensées*)

In recent years, economics has experienced an inflow of refreshing ideas thanks to the addition of elements from psychology into economic models (see Rabin (1998) and Tirole (2002) for partial surveys). The most recent literature has attempted to incorporate intrapersonal tensions in those models. The present paper provides a step in that direction.

The basic premise of our research is the existence of three types of brain conflicts. First, a conflict between the type and amount of information available in different brain areas; we call it an "asymmetric information conflict." Second, a conflict between the importance attached to immediate vs. temporally distant events; we call it a "temporal horizon conflict." Third, a conflict between the weight attached to tempting vs. not tempting goods; we call it an "incentive salience conflict." Starting from these three assumptions about the architecture of the brain, we construct an orthodox multi-period, multi-action model. The model is solved with tools adapted from mechanism design and used to explain some well-documented behavioral anomalies.

The paper can be summarized as follows. In section 2, an individual chooses a pleasant activity (consumption) and an unpleasant activity (labor) during two periods. Activities are linked through a standard intertemporal budget constraint. To model the temporal conflict, we divide the individual into an impulsive, myopic system (the "agent", he) who is interested solely in current utility and a cognitive, forward-looking system (the "principal", she) who weights equally utility at all remaining dates. The informational conflict is then incorporated by assuming that the marginal value of consumption varies from period to period and is only known to the myopic system. Last, the cognitive system has control over the impulsive system and can impose her preferred choices. Given the temporal and informational conflicts, the principal cannot reach her first-best levels of consumption and labor. Instead, she proposes a menu of pairs where the levels of both activities are linked within each period and lets the agent choose among these pairs. In other words, for the consumption and labor example, we show the endogenous emergence of a self-disciplining "work more today if you want to consume more today" intrapersonal rule of behavior (Proposition 2). The general properties of this rule hold for any amount of informational conflict and any length of temporal conflict (Proposition 3).

As developed in section 3, this result has interesting behavioral implications. First, it rationalizes narrow choice bracketing, a well-documented practice based on local (rather

than global) optimization that standard economic theories have problems explaining (Read et al., 1999). Indeed, by separating consumption into arbitrarily defined categories (clothing, entertainment, etc.) and imposing a negative relationship between expenditures in each of them, the principal achieves some discipline on the expenses incurred by the agent. This also helps understanding the simultaneous feeling of wealth and poverty described by Heath and Soll (1996). Second, our psychological personal rule can help understanding some empirical findings difficult to reconcile with the traditional life-cycle theory of consumption (Shefrin and Thaler, 1988). It predicts that consumption tracks earned income, simply because our self-disciplining rule can be more easily implemented in periods with greater access to labor. It also predicts an imperfect substitutability between mandatory and discretionary savings. Third, our theory derives a preference for the present based exclusively on the amount of information asymmetry between our two systems. The properties of this endogenously determined rate of time-preference are consistent with modern behavioral theories of choice over time: the period-to-period discount rate falls monotonically (the main characteristic of hyperbolic discounting) and discount rates are different for different categories of activities.

In section 4, we abstract from the temporal dimension and focus instead on the informational and incentive conflicts. The individual must allocate resources between a tempting good or activity whose relative desirability is only known to the agent, and a non-tempting one. We introduce the concept of drive by assuming that, for all valuations of the tempting good, the agent has a willingness to overconsume it relative to the fundamental preferences of the principal. We show that, if the conflict is increasing in the valuation, the principal sets a cap in the consumption of the tempting activity but, apart from that, she leaves full freedom to the agent (Proposition 4). This simple, non-intrusive "do what you want as long as you don't abuse" rule-of-thumb is very different from the inquisitive rule developed under temporal conflict. The choices of the agent are less likely to be constrained the smaller the conflict between brain systems and the higher the value attached by both systems to the tempting good (Proposition 5). When the conflict is decreasing in the valuation, we show that it may be optimal for the principal to waste some resources, only as a commitment device against incurring excesses.

The main justification for our informational, temporal and incentive conflicts in the brain comes from neuroscientific research. Section 1 reviews this evidence.[1] However,

[1]For clear summaries of how neuroscience can help economics, see Camerer et al. (2004, 2005).

the literatures in psychology and, to a lesser extent, economics have also addressed these issues. The remainder of this section summarizes the main findings.

Although highly controversial in economics, informational conflicts within the individual are widely documented in psychology. Some influential theories in social psychology rely on this assumption. Cognitive dissonance (Festinger, 1957) is based on the idea that an individual can simultaneously hold two contradictory beliefs. When it happens, the person acts upon one of them to reduce the discomfort created by such inconsistency. According to the theory of self-deception (Gur and Sackeim, 1979), one of these contradictory beliefs may not be subject to awareness, and this unawareness will be motivated. Last, self-perception theory (Bem, 1967) makes a stronger statement: individuals not only ignore their own attitudes, emotions and other internal states but they even make inferences about them based upon the observation of their own behavior. In other words, the individual is like an outside observer who relies on external cues to infer his inner states.[2] As for economics, Bodner and Prelec (2003) is the only existing formal study of asymmetric information within the individual. The authors focus on self-signaling, or how the "gut" who possesses some information that cannot be introspected by the "mind" uses actions to signal preferences to himself.

Temporal conflicts have also been stressed in psychology (see e.g. Ainslie (1992)). They are somewhat more accepted in economics than informational conflicts, either under hyperbolic discounting (Strotz (1956), Laibson (1997) and others) or under some other formulation of the self-control problem (Caillaud et al. (1999), Gul and Pesendorfer (2001) and others).[3] A strand of this literature has studied the effects of imperfect self-knowledge on decision-making.[4] A crucial characteristic of these studies is that the temporal and informational conflicts occur between periods (individual at date t vs. individual at date $t+1$). Instead, we stress the existence of these conflicts within each period, that is, one system of the brain "disagreeing" with another system. Hence, the analogy of the brain as a multi-agent organization. In that respect, our model is closer to Thaler and Shefrin (1981) and Shefrin and Thaler (1988) which were the first works that divided the brain into

[2]See also Bargh and Williams (2006) for a recent perspective on the effect of non-conscious mental processes on goal pursuits.

[3]See Caillaud and Jullien (2000) for a review of different ways to model time-inconsistent preferences, Caplin and Leahy (2001) for the time-inconsistency effect generated by anticipatory feelings and Bénabou and Pycia (2002) for a discussion of the link between the different approaches.

[4]See e.g. Carrillo and Mariotti (2000), Brocas and Carrillo (2004), Bénabou and Tirole (2004) and Bataglini et al. (2005).

a forward-looking and a myopic system. These studies explain the benefits of commitment devices such as mandatory pension plans and lump-sum bonus in promoting savings. They have been elegantly extended and further developed by Fudenberg and Levine (2005) and Loewenstein and O'Donoghue (2005). The first paper argues that the split-self approach can explain dynamic preference reversals and the paradox of risk-aversion in the large and in the small. The second one shows that this framework sets a parsimonious benchmark to study the optimal decision to exert willpower. None of these works, however, consider asymmetric information or incentive salience, two key driving forces behind our results.

The last of our conflicts, the biasing effect of affection on cognition, has received a growing interest among scholars across disciplines. It has been argued that the affective system helps (Damasio, 1994), constrains (Elster, 2004) or prevents (Baumeister, 2003) the cognitive system from making optimal choices. Loewenstein (1996) provides a mathematical representation of the effect of emotions (anger, fear) and drives (hunger, sexual arousal) on decision-making. The paper argues that these visceral factors cause individuals to behave contrary to their long-term best interest. This dichotomy between impulsive and reflective behavior has also been the object of recent neuroeconomic research. Benhabib and Bisin (2005) studies consumption of an individual who can invoke either a costless automatic process which is susceptible to temptation or a costly control process which is immune to temptation. Bernheim and Rangel (2004) analyze addiction under the assumption that the individual operates in either a 'cold mode' where he selects his preferred alternative or a 'hot mode' where choices may be suboptimal given preferences. Note that, in these dual-system models, information is complete. Impulsive choices are automatic responses to exogenous shocks or environmental cues. By contrast, in our model, the agent has a well-defined "goal." It just happens that his motivation is biased and, because of the superior information, he ends up affecting choices.

Before reviewing the neuroscience evidence, a clarification is in order. On the one hand, we advocate a literal interpretation of our dual-system model: the brain is, and therefore should be modelled as, a multi-agent organization with competing systems. On the other hand, the revelation games, incentive contracts and optimization processes are "as if" mechanisms, just like in standard economic problems. An advantage of the normative approach followed in the paper is that it identifies some general properties of individual behavior that can be compared to the observed choices.

1 Conflicts in the brain: some evidence from neuroscience

Brain modularity, or the idea that different areas of the brain perform different functions, is a well-accepted neurobiological fact.[5] There is also ample evidence that brain systems are often in competition and conflict (see for example the reviews by Poldrack and Rodriguez (2004) on competition between memory systems and Miller and Cohen (2001) on competition between information processing systems). Recent research suggests that some of these conflicts can even be resolved through a "mediator", the anterior cingulate cortex, which either detects and compensates for errors (Gehring et al. (1993)), or monitors competition (Carter et al. (1998), Kerns et al. (2004)).

As discussed in the previous section, the basic premises of our analysis are the existence of informational, temporal and incentive conflicts in the brain. We proceed to a brief review of the evidence in neuroscience that supports each of these conflicts as well as the connections among them.

1. Asymmetric Information. Although it has not been emphasized much in the neuroeconomics literature, asymmetric information is, for purely anatomical and evolutionary reasons, arguably the least controversial of the conflicts proposed here. Neural connectivity is a strongly limited resource that evolution spends sparingly. As a result, most brain areas are unidirectionally connected to others. These restrictions *physiologically* constrain the flow of information. Neuroscientific research provides many examples of informational asymmetries using brain imaging techniques (PET scan and fMRI). Studies have shown the activation of ventral striatum, right striatum and amygdala in response to novelty, implicit learning and fear, always without the awareness of subjects (see Berns et al. (1997), Rauch et al. (1997) and Whalen et al. (1998), respectively). Research on individuals with brain lesions reveals similar dissociations. Despite their having an intact declarative memory, patients with a damage in the neostriatum and the amygdala exhibit, respectively, an impaired ability for gradual learning and an impaired capacity to acquire conditioned responses to emotional stimuli (see Knowlton et al. (1996) and Bechara et al. (1995)).

2. Temporal horizon. The evidence of a time-evaluation conflict is more indirect, and yet more popular in neuroeconomics (Thaler and Shefrin (1981) and others). On the farsighted end, Damasio (1994) demonstrates that damage in the ventromedial prefrontal

[5] By contrast, it has been demonstrated by anatomists and neuroscientists that, contrary to the popular view based on theories developed in the 1940s and 1950s, reason and emotion *do not* pertain to two distinct brain systems (see LeDoux (1996, ch. 4) for an articulate, non-technical historical perspective).

cortex impairs the ability of patients to engage in long term planning. This severe myopia or insensitivity to future prospects is confirmed by Bechara et al. (1999) using a gambling task experiment. On the short-sighted end, LeDoux (1996) shows that the amygdala plays a crucial role in the expression of impulsive, emotional behavior. Bechara et al. (1999) elaborate also on this role of the amygdala and conclude that patients with lesions in that area have an impaired capacity to evaluate immediate gratifications. Taking both evidences together, Bechara (2005) constructs a neural theory of willpower. The author distinguishes between an impulsive system (striatum and amygdala) which processes information about immediate prospects and a reflective system (ventromedial and dorsolateral prefrontal cortex as well as anterior cingulate) which processes information about future prospects. These two broadly defined regions roughly correspond to our myopic and forward-looking systems (see Bechara (2005, Fig. 1)). McClure et al. (2004) takes the analysis one step further. Based on their fMRI experiments, the authors argue that the interaction between short-sighted and far-sighted systems provides neuroscientific support for hyperbolic discounting.

3. Incentive salience. The importance of impulses and urges in the behavior of emotional an addicted subjects has long been recognized but rarely modelled in economics. The innovative work in neuroscience by Robinson and Berridge (2003) and Berridge (2003) shows that one system mediates the feeling of pleasure and pain (the "liking" system) and a different system mediates the motivation or incentive to seek pleasure and avoid pain (the "wanting" system). Using pharmacological manipulations, the authors demonstrate that intervention in the mesolimbic dopamine system (MDS) can enhance the willingness of rats to work for food (wanting) without affecting the pleasure of eating it (liking). In a related experiment, subliminal stimuli can alter manifested choices of consumers (wanting decision) without affecting the expected pleasure derived from the commodities (liking outcome). Although, their work is particularly relevant for addiction (see Robinson and Berridge (2003) and the economic model proposed by Bernheim and Rangel (2004)), this "incentive salience" mechanism also applies to other impulse-driven choices (Berridge, 2003). The authors acknowledge that wanting and liking interact through an intricate web of brain circuits. However, they emphasize the role of the nucleus accumbens and the amygdala in the mediation of wanting, and the role of the prefrontal cortex in overriding MDS-generated impulses (Berridge and Robinson (2003, Fig. 2)). Last, it is suggested that motivational salience can be manifested without conscious awareness.

The combination of evidence about asymmetric information, temporal horizon and incentive salience provides some interesting insights. First, decisions (including motivational salience) in the absence of explicit knowledge roughly originate in the areas of the brain that we labelled as impulsive and short-sighted (ventral striatum and amygdala among others). Second, planning, mediation, anticipation of future events, and other high level cognitive functions are roughly located in the areas of the brain that we labelled as reflective and far-sighted (prefrontal cortex and anterior cingulate among others). Third, "top-down" control is exerted by the reflective system to regulate behavior whereas "bottom-up" influences emanate from the impulsive system to bias choices (Miller and Cohen (2001), Bechara (2005)). Obviously, this review constitutes only a small fraction of the current neuroscientific research on the subject. Nevertheless, taken together, it provides support for a brain architecture based on a partly uniformed, forward-looking planner and an informed, short-sighted, motivationally biased doer.

2 Temporal and informational conflicts in the brain

We consider an individual who takes two actions during a finite number of periods. Actions can be pleasant (purchasing commodities, engaging in leisure activities, etc.) or unpleasant (dieting, exercising, working, etc.). For expositional ease and without loss of generality, we will assume that actions at date t ($\in \{1, 2, \ldots, T\}$) consist in (pleasant) consumption c_t ($\geqslant 0$) and (unpleasant) labor n_t ($\in [0, \bar{n}]$). The instantaneous utility of the individual is given by the following simple equation:

$$U_t(c_t, n_t; \theta_t) = \theta_t \, u(c_t) - n_t$$

where $u' > 0$ and $u'' < 0$. The parameter θ_t captures the idea that the willingness, need or urge to consume varies from period to period. Each θ_t (sometimes referred to as "valuation" or "type") is independently drawn from the same continuous distribution in $[\underline{\theta}, \overline{\theta}]$ with $\overline{\theta} > \underline{\theta} > 0$, a strictly positive density $f(\theta_t)$ for all θ_t, and a cumulative distribution function $F(\theta_t)$ that satisfies the standard monotone hazard rate conditions ($\frac{d}{d\theta}\left[\frac{F(\theta)}{f(\theta)}\right] > 0$ and $\frac{d}{d\theta}\left[\frac{1-F(\theta)}{f(\theta)}\right] < 0$).

Our first brain conflict, namely the differences in time-horizon, is modelled in the Thaler and Shefrin (1981) tradition. First, there is one entity, the "principal" (she) who is utilitarian, cognitive and forward-looking. Second, there is, at each date t, another entity, "agent-t" (he) who is selfish, impulsive and myopic. Agent-t maximizes his instantaneous

utility $U_t(c_t, n_t; \theta_t)$ without any concern for the past or the future. The principal, on the other hand, maximizes the sum of utilities of agents in the remaining periods. This time-evaluation conflict of the self has been suggested in many disciplines. Thaler and Shefrin provide a first formalization in economics under a "Planner and Doer" label. Bechara refers to the "Reflective and Impulsive" systems in his neurocognitive theory of willpower, and identifies some key brain structures that belong to each of these systems. In this paper, we adopt a more neutral "Principal and Agent" terminology borrowed from contract theory.

In order to sharpen the contrast between principal and agent but, most importantly, to minimize the exogenous reasons for time-preference, we assume that the principal weights equally the utility of present and future agents. Formally, the intertemporal utility S_t of the principal from the perspective of date t is:

$$S_t = \sum_{s=t}^{T} U_s(c_s, n_s; \theta_s)$$

In economic terms, each agent has a discount factor $\delta = 0$ and the principal has a discount factor $\delta' = 1$. In what follows, we assume that the principal can control *at no cost* the actions to be taken at date t. However, she may decide to delegate control (within any desired limits) to agent-t or, in other words, let agent-t affect the final choice. This formalization captures the idea developed in neuroscience that the reflective system implements top-down control on choices whereas the impulsive system exerts bottom-up influence.[6]

For each unit of labor, the individual obtains one unit of income that can be consumed in any period. There is a perfect capital market where the individual can save and borrow at the exogenous, positive risk-free rate r. As a result and assuming without loss of generality no exogenous sources of income, the individual only needs to satisfy the following intertemporal budget constraint:

$$\sum_{t=1}^{T} c_t(1+r)^{T-t} \leqslant \sum_{t=1}^{T} n_t(1+r)^{T-t}$$

This formalization has an immediate but important difference with the standard life-cycle model with one decision (e.g., consumption) and an exogenous (deterministic or stochastic)

[6]For the purpose of our model, it can also be assumed that agent-t is always in charge of decisions and the principal can costlessly restrict the set of alternatives at his disposal. This is the approach followed for example by Thaler and Shefrin (1981) and Fudenberg and Levine (2005), except that restrictions are costly in their models. Note, however, that this alternative formulation where reward circuits are assumed to have control over actions has a weaker neurobiological foundation.

income stream: future consumption can be increased not only by increasing savings (i.e., reducing current consumption) but also by increasing current or future labor. In other words, there is scope for rules that "compensate" pleasant (consumption) with unpleasant (labor) activities at a given period. This will play a crucial role in the analysis.

Note, on the other hand, that neither an endogenous source of income nor a strict budget constraint are necessary ingredients for our theory. Situations where a fixed budget must be allocated between two costly, pleasant activities (say, expenditures in entertainment and clothing) can also be captured with our model (see the interpretation of the results discussed in section 3.1 and the related model proposed in section 4). Similarly, the budget constraint can be replaced by a consumption internality, where current actions affect the utility of future actions (think for example of a tasteful meal high in cholesterol that decreases future health). Overall, the only crucial ingredient is the presence of several actions that are, in one way or another, intertemporally linked.

2.1 Benchmark case: full information

As a benchmark for our analysis, consider a two-period horizon with full information. Given that the principal can impose her desired levels of consumption and labor at each period, the preferences or even the "existence" of the agents is irrelevant. The program $\mathbf{P^o}$ that the principal solves is:

$$\begin{aligned} \mathbf{P^o}: \quad \max_{\{c_1,n_1,c_2,n_2\}} \quad & \theta_1\, u(c_1) - n_1 + \theta_2\, u(c_2) - n_2 \\ \text{s.t.} \quad & c_t(\theta_t) \geqslant 0, \quad n_t(\theta_t) \in [0,\bar{n}] \quad \forall\, t,\, \theta_t && (\mathrm{F_t}) \\ & c_1(\theta_1)(1+r) + c_2(\theta_2) \leqslant n_1(\theta_1)(1+r) + n_2(\theta_2) && (\mathrm{BB}) \end{aligned}$$

where $(\mathrm{F_t})$ is a feasibility constraint on c_t and n_t and (BB) is the intertemporal budget constraint. Our first preliminary result characterizes the solution to this problem.

Proposition 1 ***(Full information)*** *The optimal pairs $(c_t^o(\theta_t), n_t^o(\theta_t))$ of consumption and labor imposed by the principal at date t to an agent-t with valuation θ_t are given by:*[7]

$$u'(c_1^o(\theta_1)) = \frac{1+r}{\theta_1} \quad \text{and} \quad n_1^o(\theta_1) = \bar{n}$$

$$u'(c_2^o(\theta_2)) = \frac{1}{\theta_2} \quad \text{and} \quad n_2^o(\theta_2) = \left(c_1^o(\theta_1) - \bar{n}\right)(1+r) + c_2^o(\theta_2)$$

[7] The proof is trivial and thus omitted. Proposition 1 implicitly assumes that $n_2^o(\theta_2) \in (0,\bar{n})$ for all θ_1 and θ_2. Sufficient conditions are $\bar{n} < \left(c_1^o(\underline{\theta})(1+r) + c_2^o(\underline{\theta})\right)/(1+r)$ and $\bar{n} > \left(c_1^o(\overline{\theta})(1+r) + c_2^o(\overline{\theta})\right)/(2+r)$. The analysis can easily be extended to other corner solutions where, for example, $n_1^o < \bar{n}$ and $n_2^o = 0$.

This proposition is straightforward. Since labor enters linearly the agents' utility function and savings have a positive net return, it is optimal for a principal who weights equally the utility of both agents to concentrate as much labor as possible in the first period. Consumption at date t is proportional to agent-t's valuation θ_t and, ceteris paribus, it is higher in period 2 than in period 1 because of the net return on savings mentioned above. As $r \to 0$, the allocation of labor between periods becomes irrelevant and inter-period differences in consumption are solely determined by differences in valuation.

In Proposition 1, consumption levels are determined only as a function of valuations and interest rates. Second period labor is then adjusted to meet the intertemporal budget constraint. In other words, there is no intra-period link between consumption and labor. Obviously, this result depends on some of our modelling assumptions (in particular, the quasi-linear utility function of agents). However, we adopt this formalization of preferences precisely because having no exogenous ties between the variables within each period constitutes an interesting benchmark of comparison.

2.2 Imperfect knowledge of impulses and desires

Our second brain conflict, namely the restriction in the flow of information, is modelled in the tradition of the contract theory literature. The principal can still impose her desired levels of consumption and labor. However, we now assume that agent-t (Pascal's heart) knows his valuation at date t whereas the principal (Pascal's reason) only knows the distribution $F(\cdot)$ from which θ_t is drawn. This captures the physiological restriction in the flow of information between different brain systems or the limited conscious awareness of motivations discussed in section 1. Alternatively, asymmetric information can also be interpreted in terms of cognitive load. The reflective system is flooded with signals. Processing them requires costly effort, a resource that must be efficiently managed. Under this competition for attention, impulsive choices are likely to influence behavior (Gilbert, 2002). Either way, differential information is problematic for the principal since her optimal decision depends on that information.

Incorporating asymmetric information in a dual-system model of the brain generates *endogenous constraints on optimal choices*. We would like to underscore the importance of this methodological contribution. As reviewed earlier, there exist other papers where the individual is split into entities that play an intra-period, non-cooperative game. However, the starting point of these studies is the existence of an exogenous cost (cost of self-control,

cost of exerting willpower, cost of attention, cost of hot choices, etc.) that inevitably leads to trade-offs (fewer resources but better allocation, costly thinking but optimal decision-making, higher current utility but increased likelihood of a future hot mode, etc.). The specific way of modelling costs crucially determines which behaviors can be rationalized. Unfortunately, it is difficult to pinpoint the right assumptions for these cost functions. Our paper proposes a different, more agnostic methodology. Rather than a *cost*, our paper rests on asymmetric information, a *constraint* on optimal decision-making. We then assume that the principal can costlessly design any mechanism she wants in order to impose on agents her favorite actions. This normative approach, borrowed from the mechanism design literature, does not presuppose a specific tradeoff. It is then difficult to anticipate which kind of deviations from optimal behavior are likely to occur. The value of the model can then be assessed by evaluating the neuroscientific foundations for informational asymmetry and the empirical relevance of the behaviors it predicts.

Under private information and assuming again a two-period horizon, the principal solves two programs. By the very nature of the problem, the principal deals with agent-1 and agent-2 sequentially, so we solve the game by backward induction. At date 2, there is no conflict of preferences between the principal and agent-2 ($S_2 \equiv U_2$). Hence, the choice set of agent-2 does not need to be restrained. Assuming that agent-1 has consumed and worked (c_1, n_1) and that the weak inequality (BB) must be satisfied, the levels of consumption and labor that agent-2 freely selects at date 2 are, just as in section 2.1:

$$u'(c_2^*(\theta_2)) = \frac{1}{\theta_2} \quad \text{and} \quad n_2^*(\theta_2) = (c_1 - n_1)(1+r) + c_2^*(\theta_2)$$

The program at date 1 is more interesting. Instead of full delegation of control, the principal now delegates it within limits by exerting some top-down control and constraining the alternatives available to agent-1. To analyze this problem, we apply familiar contract theory techniques (see e.g. Fudenberg and Tirole (1991, ch. 7)) to this unusual optimization program. More precisely, the principal delegates the choice of consumption and labor to agent-1 but restricts his options to a menu of pairs $\{(c_1(\theta_1), n_1(\theta_1))\}$. Agent-1 is free to choose any of these pairs. Applying the revelation principle, this direct mechanism achieves the maximal (second-best) welfare of the principal if it solves the following program $\mathbf{P}^*$:

$$\begin{aligned}
\mathbf{P}^* : \max_{\{(c_1(\theta_1), n_1(\theta_1))\}} \quad & S_1 = \int_{\underline{\theta}}^{\overline{\theta}} \theta_1\, u(c_1(\theta_1)) - n_1(\theta_1) + E_{\theta_2}\Big[\theta_2\, u(c_2^*(\theta_2)) - n_2^*(\theta_2)\Big]\, dF(\theta_1) \\
\text{s.t.} \quad & \theta_1\, u(c_1(\theta_1)) - n_1(\theta_1) \geqslant \theta_1\, u(c_1(\tilde{\theta}_1)) - n_1(\tilde{\theta}_1) \qquad \forall\ \theta_1,\, \tilde{\theta}_1 \qquad \text{(IC)} \\
& c_1(\theta_1) \geqslant 0, \quad n_1(\theta_1) \in [0, \bar{n}] \qquad \text{(F)}
\end{aligned}$$

In the new program, the principal maximizes expected welfare and must satisfy an incentive compatibility (IC) constraint.[8] This constraint ensures that an agent-1 with valuation θ_1 weakly prefers the menu $(c_1(\theta_1), n_1(\theta_1))$ designed for him rather than the menu $(c_1(\tilde{\theta}_1), n_1(\tilde{\theta}_1))$ designed for someone with valuation $\tilde{\theta}_1 \neq \theta_1$. Note that the constraint (BB) is binding and embedded in the second period choices $(c_2^*(\theta_2), n_2^*(\theta_2))$. The solution to program $\mathbf{P}^*$ characterizes the second-best levels of consumption and labor at date-1 from the principal's viewpoint given the information asymmetry.

Proposition 2 ***(Asymmetric information with temporal conflict)*** *The principal offers to agent-1 a menu* $\{(c_1^*(\theta_1), n_1^*(\theta_1))\}_{\theta_1=\underline{\theta}}^{\theta_1^*}$ *of consumption and labor pairs such that:*

$$u'(c_1^*(\theta_1)) = \frac{1+r}{(1+r)\theta_1 + r\left(\frac{F(\theta_1)}{f(\theta_1)}\right)}$$

$$n_1^*(\theta_1) = \bar{n} - \left[\overline{\theta}\, u(c_1^*(\overline{\theta})) - \theta_1\, u(c_1^*(\theta_1)) - \int_{\theta_1}^{\overline{\theta}} u(c_1^*(x))dx\right]$$

For every valuation $\theta_1 \in [\underline{\theta}, \theta_1^*]$*, agent-1 chooses a different pair* $(c_1^*(\theta_1), n_1^*(\theta_1))$*, where* $dc_1^*/d\theta_1 > 0$ *and* $dn_1^*/d\theta_1 > 0$*. For all valuations* $\theta_1 \in (\theta_1^*, \overline{\theta}]$ *agent-1 chooses the same pair* $(c_1^*(\theta_1^*), \bar{n})$ *as an agent-1 with valuation* θ_1^**.*[9] *The principal allows agent-2 any pair of consumption and labor provided that it satisfies* (BB)*. Agent-2 selects:*

$$u'(c_2^*(\theta_2)) = \frac{1}{\theta_2} \quad \text{and} \quad n_2^*(\theta_2) = \Big(c_1^*(\theta_1) - n_1^*(\theta_1)\Big)(1+r) + c_2^*(\theta_2)$$

The idea behind this proposition is intuitive. The utilitarian principal would like agent-1 to consume $c_1^o(\theta_1)$ and work $\bar{n}$ (see Proposition 1). However, this menu of contracts is not incentive compatible. To solve this problem, the principal could delegate the choices to the agent. In that case, the myopic and selfish agent-1 would overconsume and underwork. Another possibility would be to select the levels of consumption and labor that maximize her *expected* welfare. This would require offering a single contract $(\check{c}_1, \check{n}_1)$ such that $u'(\check{c}_1) = (1+r)/E[\theta_1]$ and $\check{n}_1 = \bar{n}$. Proposition 2 shows that the optimal plan is different from that. Intuitively, to avoid overconsumption, the principal proposes the following rule to agent-1: "Reveal your consumption needs. The higher the value you say, the higher will

[8] Contrary to most contract theory problems, this program has no participation constraint. Note, however, that the bounds $c_1 \geqslant 0$ and $n_1 \leqslant \bar{n}$ play a related role in ensuring a minimum utility to the agent. Standard techniques need to be modified to deal with this variation of the problem.

[9] See the appendix for the formal determination of the cutoff θ_1^*.

be the quantity I will allow you to consume but the higher will be the amount of work I will ask you in exchange." Demanding more work in exchange of more consumption is the best mechanism the principal can use to counter agent-1's lack of concern for the future.

Note that different valuations do not always translate into different choices. The reason for pooling lies in the absence of an exogenous minimum utility that can be secured by agent-1 (see footnote 8). As in the standard mechanism design literature, the principal could sort out agent-1's type for all θ_1. However, since labor is bounded above by $\bar{n}$, this would require too little work for low valuations and too much consumption for high valuations. The principal prefers to attenuate these two inefficiencies by granting the same consumption and requiring maximum labor for all valuations above a certain cutoff θ_1^*.

It is important to realize that the positive relation between the intertemporal levels of consumption and labor ("work more in your lifetime if you want to consume more in your lifetime" or, formally, the correlation between $c_1^* + c_2^*$ and $n_1^* + n_2^*$) *is not* a result but, instead, a consequence of the (BB) constraint. By contrast, the self-disciplining rule "work more today if you want to consume more today" *is* a main result of the dual-system model with asymmetric information. It is neither first-best nor an ad-hoc, externally imposed restriction. Instead, it emerges as the internal, self-imposed, endogenously optimal second-best rule designed by the cognitive system to counter the tendency of the impulsive system to indulge in current satisfaction. Thus, the model provides foundations for self-imposed behaviors such as: "I go to this dinner party only if I first exercise during one hour", "I watch the soccer game once I have finished the referee report", "I eat a slice of this apple pie but then I forego sugar in my coffee", etc.[10]

Last, the principal-agent literature relies either on self-enforcing contracts (reputation) or enforceability by a third party (court). A pure contracting interpretation of our model is obviously too literal. It also poses the problem of enforceability. Since both activities can hardly be undertaken simultaneously, what prevents the principal from reneging on her promise of consumption once labor has been sunk?[11] Agents' myopia impedes a dynamic self-enforcement mechanism. We hypothesize that brain areas such as the anterior cingulate which are known to act as mediators (Carter et al. (1998), Kerns et al. (2004))

[10] The unit of time is defined a bit loosely. In McClure et al. (2004), rewards are immediate if enjoyed within the day and deferred if enjoyed after 2 weeks or 4 weeks. We have in mind a similar horizon, where two activities undertaken the same day are considered within the period.

[11] We thank Douglas Bernheim for raising this issue.

may be responsible for ensuring that each system "keeps its end of the deal". Both a theoretical model and a neuro-experimental investigation of this issue would be interesting.

2.3 Enlarging the temporal horizon

A natural extension of the current analysis consists in increasing the number of periods. It is straightforward to see that setting a large but finite horizon T (> 2) does not affect qualitatively the choices selected under full information. If the maximum per-period labor $\bar{n}$ is restricted in a way that the individual finds it optimal to work in every period, the consumption granted to agent-s with $s \in \{1, 2, ..., T\}$ is given by:

$$u'(c_s^o(\theta_s)) = \frac{(1+r)^{T-s}}{\theta_s}$$

As before, labor is maximum in the first $T-1$ periods and adjusted in the last one to meet the budget constraint. Formally, $\sum_{s=1}^{T-1}(1+r)^{T-s}\ \bar{n} + n_T^o(\theta_T) = \sum_{s=1}^{T}(1+r)^{T-s}c_s^o(\theta_s)$.

When valuations are uncorrelated across dates or agents have no concern for the future, the uninformed principal does not need to worry about dynamic effects when dealing with agent-s. The behavior under asymmetric information is then similar to the two-period case. There are however a couple of technical issues that are reminiscent of the two-period analysis. First, the principal needs to make sure that sorting is possible at every period (formally, $\bar{n}$ must be such that $n_s^*(\underline{\theta}) \geqslant 0$ for all s). Second, at each period s ($< T$) there is a threshold θ_s^* such that, for all valuations above that cutoff, agent-s is granted the same consumption and required maximum labor. Assuming that $\bar{n}$ is such that the principal can induce sorting, we can determine the levels of consumption and labor at each period.

Proposition 3 ***(Extended horizon)*** *At date $s \in \{1, \ldots, T-1\}$, the principal offers to agent-s a menu $\{(c_s^*(\theta_s), n_s^*(\theta_s))\}_{\theta_s=\underline{\theta}}^{\theta_s^*}$ of consumption and labor pairs such that:*

$$u'(c_s^*(\theta_s)) = \frac{(1+r)^{T-s}}{(1+r)^{T-s}\,\theta_s + \left[(1+r)^{T-s} - 1\right]\left(\frac{F(\theta_s)}{f(\theta_s)}\right)}$$

$$n_s^*(\theta_s) = \bar{n} - \left[\overline{\theta}\,u(c_s^*(\overline{\theta})) - \theta_s\,u(c_s^*(\theta_s)) - \int_{\theta_s}^{\overline{\theta}} u(c_s^*(x))dx\right]$$

Agent-s chooses $(c_s^*(\theta_s), n_s^*(\theta_s))$ if $\theta_s < \theta_s^*$ and $(c_s^*(\theta_s^*), \bar{n})$ if $\theta_s \geqslant \theta_s^*$. Agent-$T$ is only required to satisfy (BB).

As can be immediately seen from Propositions 2 and 3, the essence of the intraperiod link between consumption and labor is preserved as the horizon is extended. However, the temporal horizon influences the levels of consumption and labor at each period. This is somewhat expected: even under full information, the number of remaining periods affects the opportunity cost of current consumption and the value of current labor. The novelty is that the amount of extra consumption that the principal needs to grant due to her lack of knowledge of the agent's desires (that is, the "informational rents") is also affected by the horizon. Since labor is directly tied to consumption, the amount of extra work also depends on T. A more in-depth analysis of the implications of this result for the dynamics of choice is deferred to section 3.3.

3 Some implication for choice over time

3.1 Choice bracketing and expense tracking

Studies in marketing and psychology show that consumers often set budgets for narrowly defined categories (clothing, entertainment, food) and track expenses against budgets (Thaler (1985), Simonson (1990)). The costs of "narrow choice bracketing" are obvious: it forces consumers to perform local rather than global maximizations. The benefits are less clear. Read et al. (1999) suggest that narrow bracketing requires less involving calculations and can be used as an effective self-disciplining mechanism to avoid excesses. However, we are not aware of any model that formalizes these advantages. The arguments seem intuitive, but not fully satisfactory. First, because nothing prevents a "broad bracketing" consumer from mimicking a "narrow bracketing" one, only when it is convenient to. Second, because the experiments of Heath and Soll (1996) demonstrate that a narrow definition of categories leads people to underconsume some goods and *overconsume others*.

We propose a different mental accounting rationale for this behavior. Following the lines of our model, suppose that an individual must intertemporally allocate a fixed income k. At each date t, most of her expenditures belong to two classes of goods: clothing ($x_t \geqslant 0$) and entertainment ($y_t \geqslant 0$). The principal can select her desired composition of expenses but ignores the agents' relative urge θ_t to consume each good. Formally, the agent's objective is $\theta_t u(x_t) + y_t$ while the principal maximizes the sum of agents' utilities under the intertemporal budget constraint $\sum_{t=1}^{T}(x_t + y_t)(1+r)^{T-t} \leqslant k(1+r)^{T-1}$.[12] If

[12] In this model, x_t plays the same role as c_t, y_t is the counterpart of $-n_t$ and the bounds on consumptions

decisions are delegated, agent-1 chooses the optimal allocation across goods in period 1 but he exhausts the budget. As demonstrated in Proposition 2, the principal can improve her welfare by explicitly separating consumption in the two categories of goods, clothing and entertainment. Then, she imposes not just a global (as obvious from the budget constraint) but also a per-period negative relationship between expenditures in each category. In this application, the simple contract takes the form "spend less on entertainment today if you want to spend more on clothing today". This strategy does not lead to first-best optimality. However, it requires a simple rule of behavior and enables the person to achieve some self-discipline, the two advantages of narrow bracketing described in the literature. Besides, if valuations for the goods are independent, the imposed negative relation will induce, on average, overconsumption of a good and underconsumption of the other. This reconciles the self-control motive for mental accounting emphasized by Thaler (1985) with the simultaneous feeling of wealth and poverty described in Heath and Soll (1996).

A similar argument can rationalize the tendency of self-employed individuals (fishermen, salesmen, writers, etc.) to work longer hours on less productive days. Consider the case of New York City cabdrivers. Assume that the forward looking principal does not observe the daily changes in the difficulty to earn money and that the myopic agent dislikes working. Delegation of control results in pernicious shirking. The principal can achieve some self-discipline and a second-best allocation of time by arbitrarily dividing the day in several subperiods (e.g., morning and afternoon). Formally, denote by l_t^m and l_t^a labor in the morning and afternoon respectively and assume that one unit of labor translates into one unit of earnings. Each day, the agent minimizes $\theta_t \psi^m(l_t^m) + \psi^a(l_t^a)$ where $\psi^m(\cdot)$ and $\psi^a(\cdot)$ represent the disutility of labor in the morning and in the afternoon, and θ_t captures the idiosyncratic shock in the relative difficulty to earn money. The principal cares about the utility of all agents and wants to meet the budget constraint $\sum_{t=1}^{T}(l_t^m + l_t^a)(1+r)^{T-t} \geqslant C \sum_{t=1}^{T}(1+r)^{T-t}$ where C represents the daily consumption of the agent. In this case, the principal proposes an incentive mechanism where labor in the afternoon is inversely related to earnings in the morning: the agent is allowed to "work less in the afternoon if (observable) earnings are high enough in the morning".[13] This can partly explain the puzzling negative elasticity of wages and hours of work documented by Camerer et al. (1997).

ensure that utility is bounded. The budget constraint is equivalent to the previous one up to a constant. The model with a detailed proof of the argument is available upon request.

[13] Again, a formal proof is available upon request.

3.2 Life-Cycle theory

The life-cycle model provides a simple framework to study intertemporal consumption. This theory makes several predictions. First, holding levels constant, the dynamics of income accumulation should not affect the dynamics of consumption. Second, the propensity to consume current income should be independent of its source. Third, if discretionary savings are positive, then an increase in pension savings should not affect total savings. Empirical analyses (e.g., Hall and Mishkin, 1982) suggest that people behave quite differently: the propensity to consume strongly depends on current income, on the source of wealth and on the level of mandatory savings (see Shefrin and Thaler (1988) and Thaler (1990) for a review of the empirical anomalies). Several theories have been proposed to explain these differences. They include bequest motives, capital market imperfections, changing preferences, self-control problems, mental accounting rules, etc. However, each departure is tailored to explain one anomaly.

Our approach may help explaining some links between income and consumption. First, it predicts that, controlling for total wealth, consumption tracks earned income: an individual consumes more when active on the labor market than when not. The intuition is simple. Assume that either the pleasure of consumption or the disutility of labor varies from period to period and is only known to the agent. The principal achieves some self-discipline with the familiar rule "work more to consume more".[14] Consumption is above its first-best level, but excesses are mitigated. By contrast, if the individual is retired or unemployed this compensatory mechanism cannot be used. To avoid maximum consumption, the principal must impose no fluctuations, that is, the consumption level that the average type selects under full information. Note that our theory predicts not only lower average levels but also smaller fluctuations in consumption during retirement or unemployment. We are not aware of any existing test of this hypothesis.

Second, according to our theory, the source of wealth affects the propensity to consume. Consumption must be granted in exchange of *costly* effort. Since the agent does not care about past or future consumption, he is willing to sacrifice any effortless income (capital gain, windfall, income borrowed against future labor, etc.) for extra consumption, independently of his type. Therefore, as income is obtained from a less costly source, the principal loses the ability of using this tool to elicit valuations. The evidence provides

[14] Obviously, "more" may not only refer to total number of hours, but also to effort, productivity, etc.

mixed support for this prediction. On the one hand, income which is more costly to obtain is spent in larger proportions: the propensity to consume regular income is greater than the propensity to consume a bonus which is itself greater than the propensity to consume a capital gain (Shefrin and Thaler, 1988). On the other hand, consumption is excessively correlated with most income changes, including windfalls. Also, the same exact gain is spent differently depending on how it is presented (e.g., cash to stockholders vs. increase in stock values). Unfortunately, our theory cannot explain these finer results.

A third and more subtle prediction relates to the effect of mandatory savings on total savings. Note that $\frac{dn_1^*(\theta_1)}{dc_1^*(\theta_1)} > 0$ and $\frac{d^2n_1^*(\theta_1)}{dc_1^*(\theta_1)^2} < 0$. This means that a higher valuation agent consumes a bigger fraction of his earned income. A mandatory savings rate (e.g., a pension plan) then constrains the choices of an agent whose valuation is above a certain cutoff $\tilde{\theta}$. Interestingly, it will also increase the savings of an agent whose valuation is below that cutoff. His consumption will remain unaffected but his labor will be increased.[15] This imperfect substitutability between mandatory and discretionary savings captures another behavioral anomaly documented in the literature (see Shefrin and Thaler (1988)).

3.3 The determinants of time-preference

From the equation that determines the first-best levels of consumption, it is immediate to notice that consumption under full information increases over time: $c_{s+1}^o(\theta) > c_s^o(\theta)$. The positive interest rate on savings implies a larger opportunity cost of consumption in early periods than in later ones. This effect is usually compensated with a preference for the present, that is, an exogenous discounting of future payoffs. It is not the case in our setting, where the principal is assumed to put equal weight in every period.

The analysis is different under asymmetric information. In order to elicit valuations, the principal grants extra consumption. The positive interest rate on savings implies that labor is more valuable in early periods. Then, the principal is willing to grant more consumption for each unit of labor in earlier than in later periods: $\frac{dc_s^*(\theta)}{dn_s^*(\theta)} > \frac{dc_{s+1}^*(\theta)}{dn_{s+1}^*(\theta)}$. As a result, consumption decreases over time: $c_s^*(\theta) > c_{s+1}^*(\theta)$. In other words, for any positive interest rate r and even if the principal puts equal weight on all periods, the informational conflict results in a positive rate of time-preference. Discounting in our model is thus derived from the brain conflict rather than assumed as an intrinsic feature of preferences.

[15] Technically, consumption is given by $\tilde{c}_1(\theta_1) = c_1^*(\tilde{\theta})$ for all $\theta_1 > \tilde{\theta}$ and $\tilde{c}_1(\theta_1) = c_1^*(\theta_1)$ for all $\theta_1 < \tilde{\theta}$. Labor is is given by $\tilde{n}_1(\theta_1) = \bar{n}$ for all $\theta_1 > \tilde{\theta}$ and $\tilde{n}_1(\theta_1) = n_1^*(\theta_1) + [\bar{n} - n_1^*(\tilde{\theta})]$ for all $\theta_1 < \tilde{\theta}$.

This conclusion can be further developed. Consider an individual with no brain conflict (or, equivalently, aware of the agents' valuations). Assume that period t ($\geqslant 2$) is, from the perspective of period 1, discounted by the following general function $\delta(t-1)$.[16] In the absence of commitment to future actions, simple calculations show that $c_s^\delta(\theta)$, her optimal consumption at date s, is:

$$u'(c_s^\delta(\theta)) = \delta(T-s)\frac{(1+r)^{T-s}}{\theta}$$

Equating this consumption to that of an asymmetrically informed principal who puts equal weight on all periods (see Proposition 3), we can identify a preference for the present or degree of impatience based exclusively on asymmetric information. Formally:

$$c_s^\delta(\theta) = c_s^*(\theta) \quad \Rightarrow \quad \delta(T-s) = \frac{\theta}{\left[(1+r)^{T-s}\right]\theta + \left[(1+r)^{T-s}-1\right]\left(\frac{F(\theta)}{f(\theta)}\right)}$$

A close scrutiny of this endogenously determined rate of time-preference reveals some intriguing properties. First and as already noted, the short-run is discounted less heavily in absolute terms than the long-run: $\delta(t+1) < \delta(t)$ (< 1) for all t. Second and more surprisingly, the short-run is discounted more heavily *in relative terms* than the long-run: $\frac{\delta(t)}{\delta(t-1)} < \frac{\delta(t+1)}{\delta(t)}$. Third, the future is discounted more heavily in activities subject to a stronger information asymmetry between the principal and the agent: as $F(\theta)/f(\theta)$ increases, both $\delta(t)$ and $\delta(t)/\delta(t-1)$ decrease.

While the first property is the most basic prediction of any study on discounting, the second and third relate to modern theories of time-evaluation. Indeed, a period-to-period discount rate that falls monotonically is the defining property of hyperbolic discounting. Although still controversial, this characteristic of discounting has received substantial support from experimental and empirical research first in psychology and now in economics (Frederick et al. (2002)). Our brain conflict hypothesis may be at the source of this behavioral anomaly.[17] As for the third property, it has been argued that individuals may not necessarily have a unique discount function (Frederick et al. (2002)). Preliminary evidence in Loewenstein et al. (2001) suggests that people exhibit different rates of time-preference for different categories of activities (e.g., repetitive tasks vs. viscerally driven behaviors). One can argue that idiosyncratic preference shocks are less predictable (and therefore

[16] Exponential discounting would correspond to $\delta(t-1) = \delta^{t-1}$.

[17] Fudenberg and Levine (2005) propose a different foundation for hyperbolic discounting and McClure et al. (2004) a neuroscientific explanation.

asymmetric information is more important) in settings subject to impulsive reactions (indulging a vice) than in recurrent tasks (flossing one's teeth). Under this assumption and other things being equal, our model predicts greater informational rents, and therefore a steeper discounting, in the former than in the latter category of activities.

4 Incentive and informational conflicts in the brain

4.1 The general setting

Temptation puts the individual in a state of mind where activities that provide a moderate objective satisfaction suddenly become irresistibly attractive. The existence of salient motivations, drives or impulsive urges has well grounded neurobiological foundations (Berridge, 2003). In this section, we incorporate our third conflict, namely the dichotomy between liking vs. wanting or reflective vs. visceral effects, in our dual-system model of the brain. To better focus on incentive salience and informational asymmetry, we abstract from the temporal conflict. At the same time, we allow for more general utility representations than before. More precisely, the individual engages in two activities, x and y, during one period. The true instantaneous payoff of the individual is:

$$U(x, y; \theta) = \theta\, u(x) + v(y)$$

where θ represents the valuation of the more tempting good x relative to the less or no tempting good y. We assume that $\theta \in [\underline{\theta}, \overline{\theta}]$ and that its c.d.f. $F(\theta)$ satisfies the same hazard rate conditions as in section 2. Function $U(\cdot)$ is the utility representation of the "liking" system (our principal), which captures how consumption of the different goods does affect welfare. However, what motivates the individual to consume is:

$$W(x, y; \theta) = \theta\, w(x) + v(y)$$

Function $W(\cdot)$ is the utility representation of the "wanting" system (our agent), which captures how perceived welfare and choices are biased by visceral influences. We assume that $u(0) = 0$, $u'(x) > 0$, $u''(x) < 0$ and $w(0) = 0$, $w'(x) > 0$, $w''(x) < 0$: both principal and agent find the good x enjoyable, although they might disagree on its marginal contribution to welfare. The two activities are bounded by the following constraint:

$$x \leqslant r(y)$$

Note that the utility of the principal and the budget constraint of the consumption and labor model studied in section 2 correspond to the special case where $v(y) = -y$ and $r(y) = y$. If, instead, we let $v'(y) > 0$ and $r(y) = b - py$, it becomes a model of allocation of consumption between two pleasurable activities (expenditures in clothing and entertainment, allocation of free time between gardening and watching TV) given a budget constraint b and prices 1 and p for goods x and y. For the rest of the section we will assume that either $v'(y) > 0$ and $r'(y) < 0$ for all y (activity y is pleasant but costly) or $v'(y) < 0$ and $r'(y) > 0$ for all y (activity y is unpleasant but generates resources). Let us call $\mathcal{U}$ and $\mathcal{W}$ the optimization programs of the principal and the agent when θ is common knowledge:

$$\mathcal{U}: \max_{x,y} \ \theta\, u(x) + v(y) \quad \text{s.t.} \quad x \leqslant r(y) \qquad \text{and} \qquad \mathcal{W}: \max_{x,y} \ \theta\, w(x) + v(y) \quad \text{s.t.} \quad x \leqslant r(y)$$

To ensure concavity of these optimization programs, we make the following assumption.

Assumption 1 *The utility of the principal and the agent satisfy:*[18]

$$\theta u''(z) + v''(r^{-1}(z))[r^{-1'}(z)]^2 + v'(r^{-1}(z))r^{-1''}(z) \leqslant 0 \quad \forall\, z,\, \theta$$
$$\theta w''(z) + v''(r^{-1}(z))[r^{-1'}(z)]^2 + v'(r^{-1}(z))r^{-1''}(z) \leqslant 0 \quad \forall\, z,\, \theta$$

Denote by $(x^F(\theta), y^F(\theta))$ and $(x^D(\theta), y^D(\theta))$ the optimal choices of principal (first-best) and agent (delegation). These pairs maximize $\mathcal{U}$ and $\mathcal{W}$, respectively:

$$\theta u'(x^F(\theta)) + v'(r^{-1}(x^F(\theta)))r^{-1'}(x^F(\theta)) = 0 \quad \text{and} \quad y^F(\theta) = r^{-1}(x^F(\theta))$$

$$\theta w'(x^D(\theta)) + v'(r^{-1}(x^D(\theta)))r^{-1'}(x^D(\theta)) = 0 \quad \text{and} \quad y^D(\theta) = r^{-1}(x^D(\theta))$$

where, in both cases, the budget constraint binds (valuable resources are never wasted). Differentiating the first-order conditions, we get $\frac{dx^F}{d\theta} > 0$ and $\frac{dx^D}{d\theta} > 0$: a higher valuation translates into a greater consumption both under first-best and under delegation. The incentive salience conflict states that the agent wants to consume an amount of the tempting good x which is considered excessive by the principal. Formally, $x^D(\theta) > x^F(\theta)$ for all θ. The following assumption ensures that this inequality holds.[19]

Assumption 2 $u'(x) < w'(x)\ \forall\, x.$

[18] These assumptions are rather weak. For instance, if $r(y)$ is linear it is sufficient to have $v''(y) \leqslant 0$.

[19] $0 = \theta\, u'(x^F) + v'(r^{-1}(x^F))r^{-1'}(x^F) \leqslant \theta\, w'(x^F) + v'(r^{-1}(x^F))r^{-1'}(x^F) \ \Rightarrow\ x^F(\theta) \leqslant x^D(\theta)$.

Given $u(0) = 0$ and $w(0) = 0$, assumption 2 also implies that $u(x) \leqslant w(x)$ for all x. Last, we denote by $T(\cdot)$ the function that transforms the utility of the agent for good x into the utility of the principal:

$$u(x) = T(w(x))$$

where $T(z) > 0$ and $T'(z) > 0$ for all z. Given assumption 2, $T'(z) < 1$ for all z.

4.2 Incentive salience and optimal delegation of choices

As in section 2, the benevolent principal maximizes welfare. Unlike before, the conflict is due to the agent being subject to urges or drives that affect his perceived utility ($W(\cdot) \neq U(\cdot)$). Under complete information, this biased motivation is irrelevant since the principal can impose her optimal pair of choices $(x^F(\theta), y^F(\theta))$. Under incomplete information, delegation results in excessive consumption of the tempting good. The principal must then design a revelation mechanism that elicits the agent's valuation. Interestingly, the options offered under incentive salience are quite different than under temporal conflict. Formally, the principal solves the following program $\mathcal{U}_{\text{AI}}$:

$$\begin{aligned}
\mathcal{U}_{\text{AI}}: \max_{\{(x(\theta),y(\theta))\}} \quad & \int_{\underline{\theta}}^{\overline{\theta}} \Big[\theta\, u(x(\theta)) + v(y(\theta))\Big] dF(\theta) \\
\text{s.t.} \quad & \theta\, w(x(\theta)) + v(y(\theta)) \geqslant \theta\, w(x(\tilde{\theta})) + v(y(\tilde{\theta})) \qquad \forall\ \theta,\, \tilde{\theta} \quad (\hat{\text{IC}}) \\
& x(\theta) \leqslant r(y(\theta)) \qquad (\hat{\text{BB}})
\end{aligned}$$

Naturally, the principal anticipates that the agent's desires to consume $(x^D(\theta), y^D(\theta))$ do not coincide with her first-best choices $(x^F(\theta), y^F(\theta))$. The solution $(\hat{x}(\theta), \hat{y}(\theta))$ to program $\mathcal{U}_{\text{AI}}$ characterizes the constrained optimum that the cognitive system can achieve given the private information and biased motivation of the affective system.

Proposition 4 ***(Asymmetric information with incentive salience)***

When $T''(z) \leqslant 0$, the principal only constrains the maximum consumption of the tempting good and requires ($\hat{\text{BB}}$). *The agent chooses his optimal level $(x^D(\theta), y^D(\theta))$ if $\theta < \hat{\theta}$ and the optimal level $(x^D(\hat{\theta}), y^D(\hat{\theta}))$ of an agent with valuation $\hat{\theta}$ if $\theta \geqslant \hat{\theta}$.*[20]

When $T''(z) > 0$, there exist n $(\geqslant 2)$ subintervals such that:
$\hat{x}(\theta) = x^D(\theta)$ *for all* $\theta \in [\underline{\theta}, \theta_1] \cap [\theta_2, \theta_3] \cap \ldots \cap [\theta_{n-2}, \theta_{n-1}]$*;*
$\hat{x}(\theta) = x^D(\theta_1)\, \forall \theta \in (\theta_1, \theta_2)$, $\hat{x}(\theta) = x^D(\theta_3)\, \forall \theta \in (\theta_3, \theta_4)$,..., $\hat{x}(\theta) = x^D(\theta_{n-1})\, \forall \theta \in (\theta_{n-1}, \overline{\theta}]$.
Moreover, resources are wasted (i.e., $x(\theta) < r(y(\theta))$) for all valuations $\theta > \theta_2$.

[20] There is also a limit case discussed in the proof where $\hat{x}(\theta) = r(\hat{y}(\theta)) = k \leqslant x^D(\underline{\theta})$ for all $\theta \in [\underline{\theta}, \overline{\theta}]$.

Contrary to Proposition 2 where optimal intervention was sophisticated and intrusive, it may now be optimal for the principal to follow a simple rule-of-thumb. Since, $u''(x) = T''(w(x))[w'(x)]^2 + T'(w(x))w''(x)$, condition $T'' \leqslant 0$ together with assumption 2 implies that $u'(x) < w'(x)$ and $u''(x) < w''(x)$: the disagreement between the principal and the agent increases with the level of consumption, and therefore with the valuation of the tempting good. In that case, the cost of letting the agent get away with his desired consumption of the tempting good is small as long as his valuation is low. When the valuation exceeds a certain threshold $\hat{\theta}$, then overconsumption becomes a serious problem and a drastic intervention in the form of a consumption cap becomes optimal. One informal way of interpreting this mechanism against temptation is the principal saying "as long as you don't abuse, you can do whatever you want." Note that the agent always make sure that the budget constraint is binding ($\hat{x}(\theta) = r(\hat{y}(\theta))$), so that resources are never wasted.

For the reader familiar with incentive theory, this form of contract should be surprising. For the sake of exposition, suppose as in section 2 that $v'(y) < 0$ (y is unpleasant) and $r(y) = y$ (so the budget constraint is $x \leqslant y$). The intuition behind the technical aspect of this result is a consequence of the three tools that the principal can use to satisfy "incentive compatibility".[21] First and trivially, the principal can let each agent choose the pair he wants. Second, she can force all agents to make the same pooling choice. Third, she can optimally select the (monotone) relation between the two variables that induces self-selection. In most problems (including the one described in Propositions 2 and 3), incentive compatibility is ensured vía the third criterion or a combination of second and third criteria when a technical property (called single-crossing) is violated. By contrast, in our setting, there is a *tension between self-selection and resource management.* On the one hand, self-selection can be induced if the relation between the pleasant and the unpleasant activity is such that $dy/dx = -\theta\, w'(x)/v'(y)$ (see appendix). On the other hand, budget balance implies that $dy/dx = 1$ (otherwise $\hat{\text{BB}}$ is not binding). Because both equalities can be satisfied for at most one type, if the principal wants to induce self-selection, she must waste resources. This is illustrated in Figure 1a: the agent consumes $x^F(\theta)$ and all types except $\tilde{\theta}$ engage in an excessive amount of y. The slanted area represents the amount of wasted resources ($x(\theta) < y(\theta)$). The other alternative for the principal is to let the agent freely select his optimal levels $\hat{x}(\theta) = x^D(\theta)$ and $\hat{y}(\theta) = x^D(\theta)$. By definition

[21] By incentive compatibility we refer to the options offered by the principal which ensure that a type-θ agent does not pick the contract designed for a type-θ' contrary to the principal's desires.

and as illustrated in Figure 1b (full line), this also results in overconsumption of the tempting good relative to first-best (dotted line) but, at least, resources are not wasted. Last, since overconsumption is especially severe for high-valuation types, the inefficiency can be limited by constraining the consumption of all agents above a certain valuation $\hat{\theta}$ (dashed line).

[Insert Figures 1a and 1b here]

This simple rule has other intuitive implications. Keeping the consumption and labor interpretation, it follows that the individual will incur excesses in both the pleasant and the unpleasant activities: the principal indulges extra consumption ($x^D(\theta) > x^F(\theta)$) but requires extra work ($y^D(\theta) > y^F(\theta)$). While self-control problems can explain overconsumption and strict rule setting can explain overwork, it is usually difficult to find reasons that explain both types of excesses at the same time (see Bénabou and Tirole (2004) for an argument based on hyperbolic discounting and imperfect recall).

One can also think of the conflict in terms of morality. The principal has a constrained willingness to engage in pleasurable activities that are socially harmful or unaccepted. The agent does not share this high-order moral disposition. Rather than imposing self-discipline for all valuations, Proposition 4 shows that the principal finds it optimal to simply limit the maximum amount of the pleasurable activity that the agent is "allowed" to enjoy (see Rabin (1995) for a different view on the effect of moral preferences and moral constraints on behavior).

Last, we can apply this mechanism to a completely different setting. Consider for instance a parent (our principal) who can constrain the options available to her offspring (our agent). The offspring privately knows the value he derives from the tempting activity, and the parent internalizes only partly his preferences. In such a situation, full delegation of choices up to a point and firm intervention thereafter is the parent's best strategy.

What happens when $T'' > 0$? The conflict between the principal and the agent can be either increasing or decreasing in consumption. In the former case, we obtain the same insights as previously: when $n = 2$, there is delegation for all $\theta \in [\underline{\theta}, \theta_1]$ and identical consumption for all $\theta \in (\theta_1, \overline{\theta}]$. More interestingly, when the conflict is decreasing or non-monotonic in valuation, the areas of delegation and strict intervention alternate. The optimal consumption path of the tempting good is illustrated in Figure 2.

[Insert Figure 2 here]

By forcing all types in an interval (say, (θ_{i-1}, θ_i)) to consume an identical amount of the tempting good x, the principal moderates the excesses. However, delegation in the next interval $[\theta_i, \theta_{i+1}]$ becomes problematic: an agent below but close to θ_i will want to pick the contract of a type-θ_i. To avoid mimicking, the principal must ensure that utility is smooth in valuation. This is achieved by imposing a significant decrease (if $v' > 0$) or increase (if $v' < 0$) of the other activity y to all agents with type θ_i and above. Since the extra restriction in y exceeds the strict needs to satisfy the budget constraint, $\hat{\text{BB}}$ is not binding anymore. Overall, the tradeoff between freedom and intervention can be summarized as follows: a longer pooling interval limits overconsumption of the tempting good but requires a bigger jump in consumption at the boundary, and therefore a larger waste of resources to ensure incentive-compatibility. Finally, note that all contractual regimes in $\mathcal{U}_{\text{AI}}$ are characterized by either full delegation or pooling, rather than self-selection as in typical mechanism design problems. The only substantial difference between the different cases is that, with $T(\cdot)$ convex, resources may not be exhausted.

We want to emphasize the fact that the incentive mechanisms under temporal salience and temptation salience are very different in nature. Under temporal salience, consumption is always excessively low from the agent's viewpoint and increasing in valuation. Under incentive salience, consumption is optimal from the agent's viewpoint for valuations below a threshold and excessively low and constant for valuations above it. The principal implements different rules simply because suboptimal choices have different costs. Under temporal conflict, excessive consumption of the tempting good has the extra cost of fewer resources being left for the future. By contrast, meeting the budget constraint is not essential since the accumulated resources can be used in the following period(s). Under incentive conflict, the opposite is true: the allocation of resources between periods is not an issue but meeting the budget constraint is crucial because unused resources are lost.

Last, from a technical viewpoint, problem $\mathcal{U}_{\text{AI}}$ shares many similarities with Amador et al. (2006), where the two activities are consumption at dates 1 and 2 and the disagreement results from hyperbolic discounting between the date-0 planner (principal) and the date-1 doer (agent). Their setting coincides with our static model with two activities under the assumption of a linear conflict (their extra preference for the present). Both papers prove the optimality of a consumption cap rule (or, equivalently, a savings threshold rule) under monotone hazard rate and linear conflict.[22] Their paper relaxes the monotone hazard

[22]See section 4.3 for an analytical characterization of this special case and some comparative statics.

rate assumption whereas our paper relaxes the conflict linearity assumption. Under either generalization (but obviously for different reasons), wasting some resources may become part of the principal's optimal strategy.

4.3 An example: linear conflict

Consider the special case of a linear conflict between the wanting and liking systems: the agent's willingness to consume the tempting good x is $w(x) = \alpha\, u(x)$ with $\alpha > 1$, so $T''(z) = 0$. Applying the methodology of Proposition 4 to this particular conflict, we obtain the following result.

Proposition 5 *Under a linear incentive conflict, choices are* $(x^D(\theta), y^D(\theta))$ *for all* $\theta \leqslant \theta_l$ *and* $(x^D(\theta_l), y^D(\theta_l))$ *for all* $\theta > \theta_l$ *where* θ_l *is such that* $\alpha\, \theta_l = E[\theta \mid \theta > \theta_l]$.

For a given valuation, the agent is less likely to make free decisions when the conflict is high and when the willingness to consume is drawn from a less favorable distribution.

Fix the utility of the principal $u(x)$. As the impulsive urges become more pronounced (α larger), the gap between the optimal choices of the principal and the motivations of the agent increases, so the former needs to control the latter more tightly in order to avoid an excessively inefficient behavior. This results in a higher probability of intervention (formally, $\partial\theta_l/\partial\alpha < 0$). It is graphically represented in Figure 3.

[Insert Figure 3 here]

Thus, for the morality or the parent/offspring interpretations, it means that more intransigent rules simply reflect stronger conflicts between the parties involved. Note that $\theta_l(\alpha) < \overline{\theta}$ for all $\alpha > 1$ and $\lim_{\alpha\to 1} \theta_l(\alpha) = \overline{\theta}$: as soon as true and perceived utility differ (even minimally), it is in the principal's best interest to intervene. Also, if the bias is sufficiently important, then the principal imposes the same action for all valuations ($\theta_l = \underline{\theta}$ for all $\alpha > E[\theta]/\underline{\theta}$). Last, one may argue that the wanting system learns over time the preferences of the liking system or that visceral impulses get to be better controlled with age and experience. In either case, if α gets closer and closer to 1 as time elapses, the incentive scheme shifts towards a more and more lenient intervention.

The distribution of valuations also affects intervention. Suppose that θ can be drawn from either $F(\theta)$ or $G(\theta)$ in $[\underline{\theta}, \overline{\theta}]$. We assume that $G(\theta)$ is more favorable than $F(\theta)$ in the sense that the functions satisfy the standard monotone likelihood ratio property

(formally $\left(\frac{g(\theta)}{f(\theta)}\right)' > 0$ for all $\theta \in [\underline{\theta}, \overline{\theta}]$). As already discussed, the optimal scheme balances the costs of overconsumption with the costs of pooling. Note that, for a given threshold θ_l, consumption is more likely to be restrictive if the distribution is more favorable. In order to avoid an excessive intervention, the principal then becomes more lenient when valuations are more likely to be high.

5 Concluding remarks

The Theory of Organizations has a long tradition in modelling the firm as a nexus of agents with incentive problems, informational asymmetries, restricted channels of communication, etc. Based on recent neuroscience research, this paper argues that individual decision-making should be studied from that same multi-agent, organizational perspective and proposes a step in that direction. A few studies have implicitly followed a similar approach before us. A main difference is that the literature has always focused on automatic processes vs. rational optimization whereas we exploit a different neuro-mechanism: the cognitive inaccessibility to our motivations.

Our model may be extended in several dimensions. We can introduce correlated valuations (or learning over time) and attenuate the conflict by assuming that agents have a positive concern for future returns. This will create a self-signaling problem different from that in Bodner and Prelec (2003) and Bénabou and Tirole (2004): agents will require extra informational rents to reveal their information since that knowledge will subsequently be used by the principal to their own detriment (the ratchet effect). We can also allow agents to invest resources that increase their productivity of labor. Technically, this will add a moral hazard stage before the contract under asymmetric information.

There are some other alleys of research that are also promising. First, other neuro-mechanisms may also achieve self-discipline. In a novel fMRI study, Zink et al. (2004) show that activity in the dorsal and ventral striatum is increased in response to monetary rewards that are contingent on subjects' behavior. This result suggests that the principal may induce the agent to work for consumption by increasing the internal pleasure derived from rewards that are earned. It would be interesting to capture this mechanism with an optimization model. Second, it would be useful to test empirically or experimentally some behavioral implications of our theory. Some results of special relevance in our model are: (i) the use of narrow choice bracketing as a self-disciplining device to overcome myopic

behavior; (ii) the lower fluctuation in consumption when the individual does not have access to labor; and (iii) the differences in discount rates for categories of activities that are subject to different degrees of idiosyncratic preference shocks.

As a final note, we would like to stress the importance of promoting collaborations between neuroscientists and economists. On one end, multiplying experiments in neuroscience will provide invaluable information to economic theorists about how to build more accurate organizational models of the brain. On the other end, developing new theories and brain models will help experimental scientists determine which hypotheses about the architecture of the brain deserve testing priority. Although it is way too early for an assessment, this methodology may eventually result in a new approach to human decision-making, moving from a decision-theory to a multi-system, game-theory formulation. We hope that our paper will modestly contribute to stimulate this line of research.

Appendix

Appendix 1. Proof of Propositions 2 and 3. The principal's objective function at date t is:

$$S_t = E_{\theta_t}\Big[\theta_t\, u(c_t(\theta_t)) - n_t(\theta_t)\Big] + \sum_{\tau=t+1}^{T} E_{\theta_\tau}\Big[\theta_\tau\, u(c^*_\tau(\theta_\tau)) - n^*_\tau(\theta_\tau)\Big]$$

where $c^*_\tau(\theta_\tau)$ and $n^*_\tau(\theta_\tau)$ are anticipated future levels. Agent-t only cares about choices at t. His utility when his valuation is θ_t and he chooses the pair $(c_t(\tilde{\theta}_t), n_t(\tilde{\theta}_t))$ is:

$$U_t(\theta_t, \tilde{\theta}_t) = \theta_t u(c_t(\tilde{\theta}_t)) - n_t(\tilde{\theta}_t)$$

Incentive Compatibility. The mechanism offered by the principal is incentive compatible if and only if $U_t(\theta_t, \theta_t) \geqslant U_t(\theta_t, \tilde{\theta}_t) \;\; \forall\, \theta_t,\, \tilde{\theta}_t$. Let $U_t(\theta_t) \equiv U_t(\theta_t, \theta_t)$. The two necessary and sufficient conditions for incentive compatibility at date t are:[23]

$$\dot{U}_t(\theta_t) = u(c_t(\theta_t)) \qquad (\mathrm{IC}_1)_\mathrm{t}$$

$$\dot{c}_t(\theta_t)\frac{\partial U_t}{\partial n_t}\left[\frac{\partial}{\partial \theta_t}\left(\frac{\partial U_t/\partial c_t}{\partial U_t/\partial n_t}\right)\right] \geqslant 0 \quad \Rightarrow \quad \dot{c}_t(\theta_t) \geqslant 0 \qquad (\mathrm{IC}_2)_\mathrm{t}$$

Feasibility. Labor $n_t(\theta_t)$ must lie in $[0, \bar{n}]$ and consumption must be positive, that is:

$$U_t(\theta_t) \geqslant \theta_t\, u(c_t(\theta_t)) - \bar{n} \equiv B^l(\theta_t) \qquad (\mathrm{FL}_1)_\mathrm{t}$$

$$U_t(\theta_t) \leqslant \theta_t\, u(c_t(\theta_t)) \equiv B^u(\theta_t) \qquad (\mathrm{FL}_2)_\mathrm{t}$$

$$c_t(\theta_t) \geqslant 0 \qquad (\mathrm{FC})_\mathrm{t}$$

Budget. At date t, the individual inherits (positive or negative) saving s_{t-1}, consumes c_t, works n_t and leaves (positive or negative) saving s_t for the next period. Since resources can a priori be thrown away, the following budget constraint inequality must hold:

$$s_{t-1}(1+r) + n_t(\theta_t) \geqslant c_t(\theta_t) + s_t \qquad (\mathrm{B})_\mathrm{t}$$

with $s_0 = 0$ (no initial resources) and $s_T \geqslant 0$ (no deficit at the end of the last period).

Program. The objective function of the principal at date t can thus be reduced to the maximization of S_t subject to $(\mathrm{IC}_1)_\mathrm{t}$, $(\mathrm{IC}_2)_\mathrm{t}$, $(\mathrm{FL}_1)_\mathrm{t}$, $(\mathrm{FL}_2)_\mathrm{t}$, $(\mathrm{FC})_\mathrm{t}$, $(\mathrm{B})_\mathrm{t}$.

Period T. There is no conflict between principal and agent-T, so $(\mathrm{IC}_1)_\mathrm{T}$ and $(\mathrm{IC}_2)_\mathrm{T}$ trivially hold. Savings at T are wasted, so $s_T = 0$. Ignoring feasibility, maximization of

[23] Techniques are standard (see e.g. Fudenberg and Tirole (1991, ch. 7)) so the proof is omitted.

S_T s.t. $(\mathrm{B})_\mathrm{T}$ implies $c_T^*(\theta_T) = c_T^o(\theta_T)$ and $n_T^*(\theta_T) = c_T^*(\theta_T) - s_{T-1}(1+r)$. We will assume that $\bar{n}$ is such that $n_T^*(\theta_T) \in [0, \bar{n}]$ for all θ_T.

No waste of resources. Given, $(c_T^*(\theta_T), n_T^*(\theta_T), s_T)$, we have that at $T-1$:

$$S_{T-1} = E_{\theta_{T-1}}\Big[\theta_{T-1}u(c_{T-1}(\theta_{T-1}))-n_{T-1}(\theta_{T-1})\Big]+E_{\theta_T}\Big[\theta_T u(c_T^*(\theta_T))-c_T^*(\theta_T)\Big]+s_{T-1}(1+r)$$

Since S_{T-1} is increasing in s_{T-1}, then $(\mathrm{B})_{\mathrm{T}-1}$ is binding. Suppose now that $(\mathrm{B})_{\mathrm{t}+1}$ to $(\mathrm{B})_{\mathrm{T}-1}$ are binding. Then, $n_T^*(\theta_T)$ can be rewritten as:

$$n_T^*(\theta_T) = c_T^*(\theta_T) + \sum_{\tau=t+1}^{T-1}(1+r)^{T-\tau}\Big(c_\tau^*(\theta_\tau) - n_\tau^*(\theta_\tau)\Big) - s_t(1+r)^{T-t}$$

Replacing into S_t, we have:

$$S_t = E_{\theta_t}\Big[\theta_t u(c_t(\theta_t)) - n_t(\theta_t)\Big] + s_t(1+r)^{T-t} + V_{t+1}$$

with:

$$V_{t+1} = \sum_{\tau=t+1}^{T} E_{\theta_\tau}\Big[\theta_\tau u(c_\tau^*(\theta_\tau)) - c_\tau^*(\theta_\tau)(1+r)^{T-\tau}\Big] + \sum_{\tau=t+1}^{T-1} E_{\theta_\tau}\Big[n_\tau^*(\theta_\tau)\Big((1+r)^{T-\tau}-1\Big)\Big]$$

Since S_t is increasing in s_t, then $(\mathrm{B})_\mathrm{t}$ is binding. Thus, we have proved that $(\mathrm{B})_{\mathrm{T}-1}$ is binding and that $(\mathrm{B})_\mathrm{t}$ is binding if $(\mathrm{B})_{\mathrm{t}+1}$ to $(\mathrm{B})_{\mathrm{T}-1}$ are binding. The combination of both results implies that $(\mathrm{B})_\mathrm{t}$ is binding for all t. In words, it is optimal not to waste resources.

Incentive compatibility and labor constraint. Given that $n_t(\theta_t) = \theta_t\, u(c_t(\theta_t)) - U_t(\theta_t)$ and that $(\mathrm{B})_\mathrm{t}$ is binding, the objective function of the principal at date t can be rewritten as:

$$S_t = E_{\theta_t}\Big[(1+r)^{T-t}\Big(\theta_t u(c_t(\theta_t))-c_t(\theta_t)\Big)+U_t(\theta_t)\Big(1-(1+r)^{T-t}\Big)\Big]+(1+r)^{T-t+1}s_{t-1}+V_{t+1}$$

which is decreasing in $U_t(\theta_t)$. Note also that, provided $(\mathrm{IC}_2)_\mathrm{t}$ is satisfied, then:

$$\dot{B}^l(\theta_t) = \dot{B}^u(\theta_t) = u(c_t(\theta_t)) + \theta_t\, u'(c_t(\theta_t))\dot{c}_t(\theta_t)) \geqslant \dot{U}_t(\theta_t) = u(c_t(\theta_t)) > 0$$

In words, the slope of the equilibrium utility is positive but smaller than the slopes of the labor feasibility constraints $B^l(\theta_t)$ and $B^u(\theta_t)$. Since we just proved that the objective function is decreasing in $U_t(\theta_t)$ (rents must be minimized), it means that $(\mathrm{IC}_1)_t$ binds at the top, that is, $U_t(\theta_t)$ binds on $B^l(\theta_t)$ at $\theta_t = \overline{\theta}$ (this, in turn, implies that $n_t(\overline{\theta}) = \bar{n}$). Let us assume that $(\mathrm{IC}_1)_t$ does not bind at any other point. Given the previous inequalities, this is true if $U_t(\underline{\theta}) < B^u(\underline{\theta})$ or, equivalently, if $n_t(\underline{\theta}) > 0$. We will neglect this inequality and check it ex-post. We then have:

$$U_t(\theta_t) = -\int_{\theta_t}^{\overline{\theta}} u(c_t(s))ds + B_l(\overline{\theta})$$

Optimal consumption. Combining the previous findings and using the standard integration by parts technique, we have:

$$S_t = E_{\theta_t}\left[(1+r)^{T-t}\Big(\theta_t\, u(c_t(\theta_t)) - c_t(\theta_t)\Big) - \Big(1-(1+r)^{T-t}\Big)u(c_t(\theta_t))\frac{F(\theta_t)}{f(\theta_t)}\right] + \Big(\overline{\theta}\, u(c_t(\overline{\theta})) - \bar{n}\Big)\Big(1-(1+r)^{T-t}\Big) + (1+r)^{T-t+1}s_{t-1} + V_{t+1}$$

So the optimal consumption maximizes S_t under $(\mathrm{IC}_2)_\mathrm{t}$ and $(\mathrm{FC})_\mathrm{t}$. Denote by $\hat{c}_t(\theta_t)$ the consumption level that maximizes the first part of the equation:

$$u'(\hat{c}_t(\theta_t))\left[(1+r)^{T-t}\,\theta_t - \big(1-(1+r)^{T-t}\big)\,\frac{F(\theta_t)}{f(\theta_t)}\right] = (1+r)^{T-t}$$

Differentiating this expression it comes that $\hat{c}_t(\theta_t)$ is increasing in θ_t. Thus, in the absence of the term $c_t(\overline{\theta})$ in S_t, $\hat{c}_t(\theta_t)$ would be the optimal consumption. Note however that by setting a consumption $\hat{c}_t(\overline{\theta})$ for an agent with valuation $\overline{\theta}$, the principal is giving rents $\overline{\theta}u(\hat{c}_t(\overline{\theta}))$ to all the agents below that valuation. In order to decrease these rents, the principal *might* prefer to constrain consumption above a certain cutoff.[24] Overall, the solution that maximizes S_t and satisfies $(\mathrm{IC}_2)_t$ has a cutoff consumption a_t such that:

$$c_t^*(\theta_t) = \begin{cases} \hat{c}_t(\theta_t) & \forall\, \theta < \theta_t^*(a_t) \\ a_t & \forall\, \theta \geqslant \theta_t^*(a_t) \end{cases}$$

where $\hat{c}_t(\theta_t^*(a_t)) = a_t$. The only remaining issue is to determine the value a_t. Three cases are possible: $a_t > \overline{a}_t \equiv \hat{c}_t(\overline{\theta})$; $a_t < \underline{a}_t \equiv \hat{c}_t(\underline{\theta})$; $a_t \in [\underline{a}_t, \overline{a}_t]$. Let:

$$\Psi_t(\theta_t, x) = \left[(1+r)^{T-t}\Big(\theta_t\, u(x) - x\Big) - \Big(1-(1+r)^{T-t}\Big)u(x)\frac{F(\theta_t)}{f(\theta_t)}\right]$$

For all $a_t > \overline{a}_t$, the welfare is:

$$\int_{\underline{\theta}}^{\overline{\theta}} \Psi_t(\theta_t, \hat{c}_t(\theta_t))\, dF(\theta_t) + \Big(\overline{\theta}\, u(a_t) - \bar{n}\Big)\Big(1-(1+r)^{T-t}\Big) + (1+r)^{T-t+1}s_{t-1} + V_{t+1}$$

This function is decreasing in a_t, so the principal always chooses $a_t \leqslant \overline{a}_t$. For all $a_t \in [\underline{a}_t, \overline{a}_t]$, the welfare of the principal in equilibrium is:

$$S_t(a_t) = \int_{\underline{\theta}}^{\theta_t^*(a_t)} \Psi_t(\theta_t, \hat{c}_t(\theta_t))dF(\theta_t) + \int_{\theta_t^*(a_t)}^{\overline{\theta}} \Psi_t(\theta_t, a_t)dF(\theta_t)$$

[24] This is a technical difference of our analysis relative to standard programs. Typically, the utility at the endpoint (where the individual rationality (IR) constraint binds) is exogenous. In our setting (with no IR constraint) the utility at the endpoint $U_t(\overline{\theta})$ is mechanism dependent, that is, it is affected by $c(\overline{\theta})$.

$$+\Big(\overline{\theta}\,u(a_t)-\bar{n}\Big)\Big(1-(1+r)^{T-t}\Big)+(1+r)^{T-t+1}s_{t-1}+V_{t+1}$$

The optimal consumption cap a_t is the one that maximizes $S_t(a_t)$. We have:

$$S_t'(a_t)=u'(a_t)K_t(a_t)-(1+r)^{T-t}\Big(1-F(\theta_t^*(a_t))\Big)\qquad\text{and}\qquad S_t''(a_t)=u''(a_t)K_t(a_t)$$

where

$$K_t(a_t)=\int_{\theta^*(a_t)}^{\overline{\theta}}\left[(1+r)^{T-t}\,\theta_t-\Big(1-(1+r)^{T-t}\Big)\frac{F(\theta_t)}{f(\theta_t)}\right]f(\theta_t)d\theta_t+\Big(1-(1+r)^{T-t}\Big)\overline{\theta}$$

Note that $K_t(a_t)$ is decreasing in a_t and that $K_t(\overline{a}_t)<0$ therefore $\overline{a}_t$ is never optimal (there is always bunching at the top). We have two cases.

If $K_t(\underline{a}_t)<0$, then $S_t'(a_t)<0$ for all $a_t\in[\underline{a}_t,\overline{a}_t]$. The optimal consumption level a_t is in $[0,\underline{a}_t]$. Therefore $c_t^*(\theta_t)=a_t$ for all $\theta_t\in[\underline{\theta},\overline{\theta}]$ and a_t maximizes:

$$\int_{\underline{\theta}}^{\overline{\theta}}\Psi_t(\theta_t,a_t)\,dF(\theta_t)+\Big(\overline{\theta}\,u(a_t)-\bar{n}\Big)\Big(1-(1+r)^{T-t}\Big)+(1+r)^{T-t+1}s_{t-1}+V_{t+1}$$

If $K_t(\underline{a}_t)>0$, there exists $\hat{a}_t\in(\underline{a}_t,\overline{a}_t)$ such that $K_t(\hat{a}_t)=0$. The welfare is strictly decreasing when $a_t\geqslant\hat{a}_t$ and it is concave when $a_t\in(\underline{a}_t,\hat{a}_t)$. If $S_t'(\underline{a}_t)<0$, we are in the same case as before (bunching for all θ_t). Last, if $S_t'(\underline{a}_t)>0$, then there exists an interior maximum $a_t^*\in(\underline{a}_t,\hat{a}_t)$ and the cutoff valuation is $\theta_t^*\equiv\theta_t^*(a_t^*)$.

Optimal labor. Given that $n_t^*(\theta_t)=\theta_t\,u(c_t^*(\theta_t))-U_t(\theta_t)$, we have:

$$n_t^*(\theta_t)=\bar{n}-\left[\overline{\theta}\,u(c_t^*(\overline{\theta}))-\theta_t\,u(c_t^*(\theta_t))-\int_{\theta_t}^{\overline{\theta}}u(c_t^*(s))ds\right]$$

In particular, for all $\theta_t\geqslant\theta_t^*(a_t^*)$, there is bunching and $n_t^*(\theta_t)=\bar{n}$. Also,

$$\frac{dn_t^*}{d\theta_t}=\theta_t\,u'(c_t^*(\theta_t))\,\frac{dc_t^*}{d\theta_t}$$

which is strictly positive for all $\theta_t<\theta_t^*$. Last, the neglected inequality $n_t^*(\underline{\theta})>0$ is automatically satisfied if $\bar{n}$ is "sufficiently large" or, more specifically, if:

$$\bar{n}>\overline{\theta}\,u(c_t^*(\overline{\theta}))-\underline{\theta}\,u(c_t^*(\underline{\theta}))-\int_{\underline{\theta}}^{\overline{\theta}}u(c_t^*(s))ds$$

Appendix 2. Proof of Propositions 4 and 5. Let $W(\theta) = \theta\, w(x(\theta)) + v(y(\theta))$. Using standard techniques (proof omitted), the incentive compatibility constraints ($\hat{\text{IC}}$) in program $\mathcal{U}_{\text{AI}}$ are equivalent to the following first- and second-order conditions:

$$\dot{W}(\theta) = w(x(\theta)) \qquad \text{and} \qquad \dot{x}(\theta) \geqslant 0$$

Also, when $v' > 0$ and $r' < 0$ or when $v' < 0$ and $r' > 0$, ($\hat{\text{BB}}$) can be rewritten as:

$$W(\theta) \leqslant \theta\, w(x(\theta)) + v(r^{-1}(x(\theta)))$$

Since $v(y(\theta)) = W(\theta) - \theta w(x(\theta))$, program $\mathcal{U}_{\text{AI}}$ can thus be rewritten as:

$$\begin{aligned}
\mathcal{U}_{\text{AI}}: \max_{\{(x(\theta),W(\theta))\}} \quad & \int_{\underline{\theta}}^{\overline{\theta}} \Big[\theta\, u(x(\theta)) - \theta\, w(x(\theta)) + W(\theta)\Big]\, dF(\theta) \\
\text{s.t.} \quad & \dot{W}(\theta) = w(x(\theta)) && (\hat{\text{IC}}_1) \\
& \dot{x}(\theta) \geqslant 0 && (\hat{\text{IC}}_2) \\
& W(\theta) \leqslant B(\theta) = \theta\, w(x(\theta)) + v(r^{-1}(x(\theta))) && (\hat{\text{BB}})
\end{aligned}$$

The equilibrium utility increases at rate $\dot{W}(\theta) = w(x(\theta))$ and the upper bound of ($\hat{\text{BB}}$) increases at rate $\dot{B}(\theta) \equiv \dot{x}(\theta)\Big[\theta w'(x(\theta)) + v'(r^{-1}(x(\theta)))r^{-1\prime}(x(\theta))\Big] + w(x(\theta))$. Given ($\hat{\text{IC}}_2$), assumption 1 and the definition of $x^D(\theta)$ as the maximum in $\mathcal{W}$, then in equilibrium:

$$\dot{W}(\theta) \lesseqgtr \dot{B}(\theta) \;\Leftrightarrow\; x(\theta) \lesseqgtr x^D(\theta).$$

Since $x^F(\theta) < x^D(\theta)$, then $(x(\theta'), y(\theta'))$ with $x(\theta') > x^D(\theta')$, yields lower utility to the principal than $(x^D(\theta'), r^{-1}(x^D(\theta')))$, provided the latter is incentive compatible at θ'. The indifference curves of the principal satisfy $x'(y) = -v'(y)/\theta u'(x)$. They are decreasing and convex if $v' > 0$ and $r' < 0$ and increasing and convex if $v' < 0$ and $r' > 0$. To satisfy incentive compatibility, $dx/dy = -v'(y)/\theta w'(x)$. Assume now that the contract entails $(x^D(\theta'), y^{ic}(\theta'))$ for some θ' with $x^D(\theta') < r(y^{ic}(\theta'))$. Consider a deviation to $x(\theta') > x^D(\theta')$ and let $y(\theta')$ be such that $(x(\theta'), y(\theta'))$ is incentive compatible. Given the previous properties, $\theta\, u(x^D(\theta')) + v(y^{ic}(\theta')) > \theta\, u(x(\theta')) + v(y(\theta'))$. This proves that it is never optimal to set $x(\theta) > x^D(\theta)$ for any θ. Therefore, from now on, we shall restrict the attention to solutions of the form $x(\theta) \leqslant x^D(\theta)$ for all θ.

Note that $W(\theta)$ enters positively in the principal's objective function. Also, $x(\theta) \leqslant x^D(\theta)$ implies $\dot{W}(\theta) \leqslant \dot{B}(\theta)$. Combining both arguments, $W(\theta)$ binds in ($\hat{\text{BB}}$) at the lower bound $\underline{\theta}$. Using ($\hat{\text{IC}}_1$) and ($\hat{\text{BB}}$), we then have:

$$W(\theta) = \int_{\underline{\theta}}^{\theta} w(x(s))ds + W(\underline{\theta}) \quad \text{with} \quad W(\underline{\theta}) = \underline{\theta} w(x(\underline{\theta})) + v(r^{-1}(x(\underline{\theta})))$$

Using standard integration by parts techniques, the problem becomes:

$$\begin{array}{lll} \mathcal{U}_{\text{AI}}: \quad \max\limits_{\{x(\theta)\}} & \displaystyle\int_{\underline{\theta}}^{\overline{\theta}} \left[\theta\, u(x(\theta)) - \theta\, w(x(\theta)) + w(x(\theta))\frac{1-F(\theta)}{f(\theta)}\right] dF(\theta) + W(\underline{\theta}) & \\ \text{s.t.} & \dot{x}(\theta) \geqslant 0 & (\hat{\text{IC}}_2) \\ & W(\underline{\theta}) = \underline{\theta} w(x(\underline{\theta})) + v(r^{-1}(x(\underline{\theta}))) & (\text{E}) \\ & x(\theta) \leqslant x^D(\theta) & (\text{D}) \end{array}$$

where (E) is the utility at $\underline{\theta}$ and (D) is the restriction on the domain. The rest of the proof proceeds as follows. First, we ignore ($\hat{\text{IC}}_2$) and (E) and find the solutions that satisfy (D). Second, we construct the solutions that also satisfy ($\hat{\text{IC}}_2$). Last, we introduce (E). Let:

$$\Lambda(x,\theta) = \theta\, T(w(x)) - \theta\, w(x) + w(x)\frac{1-F(\theta)}{f(\theta)}.$$

where $\Lambda(0,\theta) = 0$; $\frac{\partial \Lambda(x,\theta)}{\partial x} = w'(x)\Big[\theta T'(w(x)) - \theta + \frac{1-F(\theta)}{f(\theta)}\Big]$; $\frac{\partial \Lambda(x,\theta)}{\partial x}\Big|_{\overline{\theta}} \leqslant 0$; $\frac{\partial^2 \Lambda(x,\theta)}{\partial x \partial \theta} = w'(x)\Big[T'(w(x)) - 1 + \Big(\frac{1-F(\theta)}{f(\theta)}\Big)'\Big] \leqslant 0$; $\frac{\partial^2 \Lambda(x,\theta)}{\partial x^2} = \frac{w''(x)}{w'(x)}\frac{\partial \Lambda(x,\theta)}{\partial x} + [w'(x)]^2\, \theta\, T''(w(x))$. Denote by $\tilde{x}(\theta)$ the interior optimum of $\Lambda(x,\theta)$, if it exists. We shall consider two different cases.

Case 1: $T''(\cdot) > 0$. $\frac{\partial \Lambda(\tilde{x}(\theta),\theta)}{\partial x} = 0$ implies $\frac{\partial^2 \Lambda(\tilde{x}(\theta),\theta)}{\partial x^2} > 0$, so $\tilde{x}(\theta)$ is the unique minimum of $\Lambda(x,\theta)$. The maxima are the corner solutions 0 or $x^D(\theta)$. Also, there exists $\tilde{\theta}$ such that for all $\theta > \tilde{\theta}$, $\Lambda(x,\theta)$ is strictly decreasing in x and the maximum is 0. For $\theta \leqslant \tilde{\theta}$, the maximum alternates between 0 and $x^D(\theta)$.

Case 1a. Suppose that the maximum at $\underline{\theta}$ is $x^D(\underline{\theta})$. Then, there exists a series of cutoffs $(\theta_0, \ldots, \theta_{2t-1}, \theta_{2t})$ where $\theta_0 = \underline{\theta}$, $\theta_{2t-1} = \tilde{\theta}$ and $\theta_{2t} = \overline{\theta}$, such that:

$$\tilde{x}(\theta) = \begin{cases} x^D(\theta) & \text{if} \quad \theta \in [\theta_s, \theta_{s+1}] \\ 0 & \text{if} \quad \theta \in (\theta_{s+1}, \theta_{s+2}) \end{cases} \qquad \forall\ s \in \{0, 2, \ldots, 2t-2\}$$

Note that $\tilde{x}(\theta)$ does not satisfy ($\hat{\text{IC}}_2$) in the neighborhood of θ_{s+1}. When adding this constraint, we could set consumption at $x^D(\theta_{s+1})$ for all $\theta \in (\theta_{s+1}, \theta_{s+2})$ (it is obviously suboptimal to go above). It may however, be preferable to start pooling at $\theta'_{s+1} < \theta_{s+1}$: the cost of $x^D(\theta'_{s+1}) < x^D(\theta)\ \forall\, \theta \in (\theta'_{s+1}, \theta_{s+1}]$ may be offset by the benefits of $x^D(\theta'_{s+1}) < x^D(\theta_{s+1})\ \forall\, \theta \in (\theta_{s+1}, \theta_{s+2})$.[25] Overall, there will exist new cutoffs $\theta'_{s+1} \in [\theta_s, \theta_{s+1}]$ such that the solution that maximizes the principal's objective under (D) and ($\hat{\text{IC}}_2$) is:

$$x^*(\theta) = \begin{cases} x^D(\theta) & \text{if} \quad \theta \in [\theta_s, \theta'_{s+1}] \\ x^D(\theta'_{s+1}) & \text{if} \quad \theta \in (\theta'_{s+1}, \theta_{s+2}) \end{cases} \qquad \forall\ s \in \{0, 2, \ldots, 2t-2\}$$

[25] The argument is the same as to where bunching should start in standard mechanism design problems when $\dot{x}(\theta) \geqslant 0$ is not automatically satisfied.

Last, let a be the optimal consumption at $\underline{\theta}$, where $a \leqslant x^D(\underline{\theta})$ to satisfy (D). Denote by $\hat{x}(\theta)$ the optimal solution of the principal's program under ($\hat{\text{IC}}_2$), (E), (D). We have $\hat{x}(\underline{\theta}) = a$ and $\hat{x}(\theta) = x^*(\theta)\ \forall\, \theta > \underline{\theta}$. The equilibrium utility of the principal is then:

$$\int_{\underline{\theta}}^{\overline{\theta}} \Lambda(x^*(\theta), \theta) dF(\theta) + \underline{\theta} w(a) + v(r^{-1}(a))$$

This utility is increasing in a, so $a = x^D(\underline{\theta})$. Overall, the optimal solution is:

$$\hat{x}(\theta) = x^*(\theta) = \begin{cases} x^D(\theta) & \text{if} \quad \theta \in [\theta_s, \theta'_{s+1}] \\ x^D(\theta'_{s+1}) & \text{if} \quad \theta \in (\theta'_{s+1}, \theta_{s+2}) \end{cases} \qquad \forall\ s \in \{0, 2, \ldots, 2t-2\}$$

It remains to determine $\hat{y}(\theta)$. The agent's utility under delegation is:

$$\begin{aligned} W^D(\theta) &= \theta\, w(x^D(\theta)) + v(r^{-1}(x^D(\theta)) && (1) \\ &= \int_{\underline{\theta}}^{\theta} w(x^D(c))dc + \underline{\theta}\, w(x^D(\underline{\theta})) + v(r^{-1}(x^D(\underline{\theta}))) && (2) \end{aligned}$$

since $\dot{W}^D(\theta) = w(x^D(\theta))$. The agent's utility under the optimal contract $(\hat{x}(\theta), \hat{y}(\theta))$ is:

$$\begin{aligned} W(\theta) &= \theta\, w(\hat{x}(\theta)) + v(\hat{y}(\theta)) && (3) \\ &= \int_{\underline{\theta}}^{\theta} w(\hat{x}(c))dc + \underline{\theta}\, w(\hat{x}(\underline{\theta})) + v(r^{-1}(\hat{x}(\underline{\theta}))) && (4) \end{aligned}$$

For all $\theta \in [\underline{\theta}, \theta'_1]$, we have $\hat{x}(\theta) = x^D(\theta)$ and $W(\theta) = W^D(\theta)$, so $v(\hat{y}(\theta)) = v(r^{-1}(x^D(\theta)))$, and resources are not wasted. For all $\theta \in (\theta'_1, \theta_2)$, we have $\hat{x}(\theta) = x^D(\theta'_1)$. Using (2) and (4), we have $W(\theta) = W^D(\theta'_1) + (\theta - \theta'_1) w(x^D(\theta'_1))$. Using (1) and (3), we get $v(\hat{y}(\theta)) = v(r^{-1}(x^D(\theta'_1)))$ and, again, resources are not wasted. For all $\theta \in [\theta_2, \theta'_3]$, we have $\hat{x}(\theta) = x^D(\theta)$ but $W(\theta) < W^D(\theta)$. Then, $v(\hat{y}(\theta)) < v(r^{-1}(x^D(\theta)))$, that is, for all $\theta \geqslant \theta_2$ resources are wasted.

Case 1b. Suppose that the maximum at $\underline{\theta}$ is 0. Following the analogous reasoning as in case 1a, the maximization of the principal's objective under ($\hat{\text{IC}}_2$) and (D) yields:

$$x^*(\theta) = \begin{cases} 0 & \text{if} \quad \theta \in [\underline{\theta}, \theta_1) \\ x^D(\theta) & \text{if} \quad \theta \in [\theta_s, \theta'_{s+1}] \qquad \forall\ s \in \{1, 3, \ldots, 2t-1\} \\ x^D(\theta'_{s+1}) & \text{if} \quad \theta \in (\theta'_{s+1}, \theta_{s+2}) \quad \forall\ s \in \{1, 3, \ldots, 2t-1\} \end{cases}$$

Adding constraint (E) to the program, modifies the solution into $\hat{x}(\theta) = a$ for all $\theta \in [\underline{\theta}, \theta_1)$ and $\hat{x}(\theta) = x^*(\theta)$ for all $\theta \in [\theta_1, \overline{\theta}]$, with $a \in [0, x^D(\underline{\theta})]$. The principal's utility is then:

$$\int_{\underline{\theta}}^{\theta_1} \Lambda(a, \theta)\, dF(\theta) + \int_{\theta_1}^{\overline{\theta}} \Lambda(x^*(\theta), \theta)\, dF(\theta) + \underline{\theta}\, w(a) + v(r^{-1}(a))$$

Let $\hat{a} = \text{argmax}_{a\in[0,x^D(\underline{\theta})]} \int_{\underline{\theta}}^{\theta_1} \Lambda(a,\theta)dF(\theta) + \underline{\theta}\, w(a) + v(r^{-1}(a))$. The optimal solution is:

$$\hat{x}(\theta) = \begin{cases} \hat{a} & \text{if} \quad \theta \in [\underline{\theta}, \theta_1) \\ x^D(\theta) & \text{if} \quad \theta \in [\theta_s, \theta'_{s+1}] \quad \forall\ s \in \{1,3,\ldots,2t-1\} \\ x^D(\theta'_{s+1}) & \text{if} \quad \theta \in (\theta'_{s+1}, \theta_{s+2}) \quad \forall\ s \in \{1,3,\ldots,2t-1\} \end{cases}$$

Using the same method as in case 1a, we can compute $\hat{y}(\theta)$. For all $\theta < \theta_1$, $\hat{x}(\theta) = \hat{a} \leqslant x^D(\theta)$ and $W(\theta) = \int_{\underline{\theta}}^{\theta} w(\hat{a})ds + \underline{\theta}\, w(\hat{a}) + v(r^{-1}(\hat{a}))$. Combining it with (3), we get that $v(\hat{y}(\theta)) = v(r^{-1}(\hat{a}))$, so resources are not wasted. For all $\theta \in [\theta_1, \theta'_2]$, consumption is $x^D(\theta)$ and, using (2) and (4), we have $W(\theta) < W^D(\theta)$. Therefore, $v(\hat{y}(\theta)) < v(r^{-1}(x^D(\theta)))$ and resources are wasted for all $\theta \geqslant \theta_1$.

Case 2: $T''(\cdot) \leqslant 0$. If $\tilde{x}(\theta)$ exists, it is the unique interior maximum. However, it is decreasing in θ so it does not satisfy ($\hat{\text{IC}}_2$). Again, there exists $\tilde{\theta}$ such that for all $\theta > \tilde{\theta}$, $\Lambda(x,\theta)$ is strictly decreasing in x, so the maximum is 0. For all $\theta \leqslant \tilde{\theta}$, $\tilde{x}(\theta)$ exists. The maximum of $\Lambda(x,\theta)$ under (D), is $\tilde{x}(\theta)$ if $\tilde{x}(\theta) \leqslant x^D(\theta)$ and $x^D(\theta)$ if $\tilde{x}(\theta) \geqslant x^D(\theta)$.

Case 2a. Since $\frac{dx^D(\theta)}{d\theta} > 0$ and $\frac{d\tilde{x}(\theta)}{d\theta} < 0$, if $x^D(\underline{\theta}) \leqslant \tilde{x}(\underline{\theta})$, then there exists θ' such that $x^D(\theta) < \tilde{x}(\theta)$ for all $\theta < \theta'$ and $x^D(\theta) \geqslant \tilde{x}(\theta)$ for all $\theta \geqslant \theta'$. To satisfy ($\hat{\text{IC}}_2$), the principal could set $x^D(\theta)$ for all $\theta < \theta'$ and $x^D(\theta')$ for all $\theta \geqslant \theta'$. However, using the same logic as in case 1a, there will exist a cutoff $\hat{\theta} \in [\underline{\theta}, \theta')$ such that (see later for its determination):

$$x^*(\theta) = \begin{cases} x^D(\theta) & \text{if} \quad \theta \in [\underline{\theta}, \hat{\theta}) \\ x^D(\hat{\theta}) & \text{if} \quad \theta \in [\hat{\theta}, \overline{\theta}] \end{cases}$$

Adding constraint (E) to the program modifies the solution into $\hat{x}(\underline{\theta}) = a$ and $\hat{x}(\theta) = x^*(\theta)$ for all $\theta \in (\underline{\theta}, \overline{\theta}]$, with $a \in [0, x^D(\underline{\theta})]$. The principal's equilibrium utility is then:

$$\int_{\underline{\theta}}^{\overline{\theta}} \Lambda(x^*(\theta),\theta)dF(\theta) + \underline{\theta}\, w(a) + v(r^{-1}(a))$$

which is increasing in a, so $a = x^D(\underline{\theta})$. Overall, the optimal solution is:

$$\hat{x}(\theta) = \begin{cases} x^D(\theta) & \text{if} \quad \theta \in [\underline{\theta}, \hat{\theta}) \\ x^D(\hat{\theta}) & \text{if} \quad \theta \in [\hat{\theta}, \overline{\theta}] \end{cases}$$

Using the same reasoning as in case 1a, resources are never wasted. Last and for the sake of completeness, we characterize $\hat{\theta}$. Given $\hat{\theta}$, the equilibrium utility of the principal is:

$$\hat{U} = \int_{\underline{\theta}}^{\hat{\theta}} \Lambda(x^D(\theta),\theta)dF(\theta) + \int_{\hat{\theta}}^{\overline{\theta}} \Lambda(x^D(\hat{\theta}),\theta)dF(\theta) + \underline{\theta}\, w(x^D(\underline{\theta})) + v(r^{-1}(x^D(\underline{\theta})))$$

The first-order condition that determines the optimal cutoff $\hat{\theta}$ is then given by (note that we would need to impose further restrictions to ensure uniqueness):

$$\frac{d\hat{U}}{d\hat{\theta}} = 0 \;\Rightarrow\; \int_{\hat{\theta}}^{\overline{\theta}} \frac{\partial \Lambda}{\partial x}(x^D(\hat{\theta}), \theta)\, dF(\theta) = 0$$

Since $\left.\dfrac{d\hat{U}}{d\hat{\theta}}\right|_{\hat{\theta}=\theta'} = \displaystyle\int_{\theta'}^{\overline{\theta}} \frac{\partial \Lambda(x^D(\theta'), \theta)}{\partial x} \frac{\partial x^D(\theta')}{\partial \theta}\, dF(\theta) < 0$, we then have that $\hat{\theta} < \theta'$.

Case 2b. Since $\frac{d\tilde{x}(\theta)}{d\theta} < 0$, if $x^D(\underline{\theta}) > \tilde{x}(\underline{\theta})$, then it is optimal to set the same consumption level for all θ. This amount is given by:

$$\hat{x}(\theta) = \hat{a} \quad \forall\, \theta \in [\underline{\theta}, \overline{\theta}] \quad \text{where} \;\; \hat{a} = \arg\max_a \int_{\underline{\theta}}^{\overline{\theta}} \Lambda(a, \theta) dF(\theta) + \underline{\theta}\, w(a) + v(r^{-1}(a))$$

Note that $\frac{\partial \Lambda}{\partial a}(x^D(\underline{\theta}), \theta) < 0$ and $\frac{d}{da}\Big[\underline{\theta}\, w(x^D(\underline{\theta})) + v(r^{-1}(x^D(\underline{\theta})))\Big] = 0$, so $\hat{a} < x^D(\underline{\theta})$.

Case 3: Special case $T''(\cdot) = 0$. Assume $w(x) = \alpha u(x)$ with $\alpha > 1$. We have:

$$\Lambda(x, \theta) = w(x)\, K(\theta) \qquad \text{where} \;\; K(\theta) = \theta \frac{1}{\alpha} - \theta + \frac{1 - F(\theta)}{f(\theta)}$$

Following the same reasoning as in case 2a, we have $\hat{x}(\theta) = x^D(\theta)$ if $\theta \in [\underline{\theta}, \theta_l)$ and $\hat{x}(\theta) = x^D(\theta_l)$ if $\theta \in [\theta_l, \overline{\theta}]$, where the cutoff θ_l is determined by the following equality:

$$\frac{d\hat{U}}{d\theta_l} = 0 \;\Rightarrow\; w'(x^D(\theta_l)) \frac{dx^D(\theta_l)}{d\theta} \int_{\theta_l}^{\overline{\theta}} K(\theta) f(\theta)\, d\theta = 0 \Rightarrow \int_{\theta_l}^{\overline{\theta}} \left[(\frac{1}{\alpha} - 1)\theta f(\theta) + 1 - F(\theta) \right] d\theta = 0 \quad (5)$$

$$\Rightarrow\; E[\theta \mid \theta > \theta_l] \;=\; \alpha\, \theta_l \quad (6)$$

Note that the cutoff θ_l is indeed a unique maximum:

$$\frac{d^2\hat{U}}{d\theta_l^2} = \frac{d}{d\theta_l}\left[w'(x^D(\theta_l)) \frac{dx^D(\theta_l)}{d\theta} \right] \int_{\theta_l}^{\overline{\theta}} K(\theta) f(\theta)\, d\theta - w'(x^D(\theta_l)) \frac{dx^D(\theta_l)}{d\theta} K(\theta_l) f(\theta_l) < 0$$

where the first term is equal to zero by (5) and $K(\theta_l) > 0$ by (5) and $dK/d\theta < 0$. Also, every type consumes the same amount ($\theta_l = \underline{\theta}$) if and only if $\left.\frac{d\hat{U}}{d\theta_l}\right|_{\theta_l = \underline{\theta}} \leqslant 0 \;\Rightarrow \alpha > \overline{\alpha} \equiv \frac{E[\theta]}{\underline{\theta}}$.

Differentiating (5): $-K(\theta_l(\alpha), \alpha) f(\theta_l(\alpha)) \dfrac{d\theta_l}{d\alpha} + \displaystyle\int_{\theta_l}^{\overline{\theta}} \frac{\partial K(\theta, \alpha)}{\partial \alpha} f(\theta) d\theta = 0 \;\Rightarrow\; \dfrac{d\theta_l}{d\alpha} < 0.$

Last, if $\left(\frac{g(\theta)}{f(\theta)}\right)' > 0$ for all θ, then $E_{G(\theta)}[\theta \mid \theta > \theta_l] > E_{F(\theta)}[\theta \mid \theta > \theta_l]$. As a result and given (6), $\theta_l^G > \theta_l^F$ where θ_l^G is the cutoff under distribution $G(\theta)$ and θ_l^F is the cutoff under distribution $F(\theta)$.

References

1. Ainslie, G. (1992), *Picoeconomics*, Cambridge University Press.

2. Amador, M., Werning, I. and G.M. Angeletos (2006), "Commitment vs. Flexibility", *Econometrica*, 74(2), 365-396.

3. Bargh, J.A. and E.L. Williams (2006), "The Automaticity of Social Life", *Current Directions in Psychological Science*, 15(1), 1-4.

4. Bataglini, M., Bénabou, R. and J. Tirole (2005), "Self-Control in Peer Groups" , *Journal of Economic Theory*, 112 (4), 848-887.

5. Baumeister, R. (2003), "The Psychology of Irrationality: Why People Make Foolish, Self-Defeating Choices", in I. Brocas and J. Carrillo *The Psychology of Economic Decisions. Vol.1: Rationality and Well-Being*, 3-16, Oxford: Oxford University Press.

6. Bechara, A. (2005) "Decision Making, Impulse Control and Loss of Willpower to Resist Drugs: a Neurocognitive Perspective", *Nature Neuroscience*, 8(11), 1458-1463.

7. Bechara, A., Damasio, H., Damasio, A., and G. Lee (1999), "Different Contributions of the Human Amygdala and Ventromedial Prefrontal Cortex to Decision-Making", *Journal of Neuroscience*, 19(13), 5473-5481.

8. Bechara, A., Tranel, D., Damasio, H., Adolphs, R., Rockland, C. and A. Damasio (1995), "Double Dissociation of Conditioning and Declarative Knowledge Relative to the Amygdala and Hippocampus in Humans", *Science*, 269, 1115-1118.

9. Bem, D.J. (1967), "Self-Perception: an Alternative Interpretation of Cognitive Dissonance Phenomena", *Psychological Review*, 74, 183-200.

10. Bénabou, R. and M. Pycia (2002), "Dynamic Inconsistency and Self-Control: A Planner-Doer Interpretation", *Economics Letters*, 77, 419-424.

11. Bénabou, R. and J. Tirole (2004), "Willpower and Personal Rules", *Journal of Political Economy*, 112, 848-887.

12. Benhabib, J. and A. Bisin (2005), "Modeling Internal Commitment Mechanisms and Self-Control: a Neuroeconomics Approach to Consumption-Saving Decisions", *Games and Economic Behavior*, 52(2), 460-492.

13. Bernheim, B.D. and A. Rangel (2004), "Addiction and Cue-Triggered Decision Processes", *American Economic Review*, 94(5), 1558-1590.

14. Berns, G, Cohen, D. and M. Mintun (1997), "Brain Regions Responsive to Novelty in the Absence of Awareness", *Science*, 276, 1272-1275.

15. Berridge, K. (2003), "Irrational Pursuit: Hyper-Incentives from a Visceral Brain", in I. Brocas and J. Carrillo eds. *The Psychology of Economic Decisions. Vol.1: Rationality and Well-Being*, 17-40, Oxford: Oxford University Press.

16. Berridge, K. and T. Robinson (2003), "Parsing Reward", *Trends in Neurosciences*, 26(9), 507-513.

17. Bodner, R. and D. Prelec (2003), "Self-Signaling and Diagnostic Utility in Everyday Decision Making" in I. Brocas and J. Carrillo *The Psychology of Economic Decisions. Vol.1: Rationality and Well-Being*, 105-126, Oxford: Oxford University Press.

18. Brocas, I. and J.D. Carrillo (2004), "Entrepreneurial Boldness and Excessive Investment", *Journal of Economics & Management Strategy*, 13, 321-50.

19. Caillaud, B. and B. Jullien (2000), "Modelling Time-Inconsistent Preferences", *European Economic Review*, 44, 1116-1124.

20. Caillaud, B., Cohen, D. and B. Jullien (1999), "Towards a Theory of Self-Restraint", *mimeo*, CERAS and Toulouse.

21. Camerer, C., Babcock, L., Loewenstein, G. and R. Thaler (1997), "Labor Supply of New York City Cabdrivers: One Day at a Time", *Quarterly Journal of Economics*, 112, 407-441.

22. Camerer, C., Loewenstein, G. and D. Prelec (2004), "Neuroeconomics: Why Economics Needs Brains", *Scandinavian Journal of Economics*, 106(3), 555-579.

23. Camerer, C., Loewenstein, G. and D. Prelec (2005), "Neuroeconomics: How Neuroscience can Inform Economics", *Journal of Economic Literature*, 43, 9-64.

24. Caplin, A. and J. Leahy (2001), "Psychological Expected Utility Theory and Anticipatory Feelings", *Quarterly Journal of Economics*, 116, 55-80.

25. Carrillo, J.D., and T. Mariotti (2000), "Strategic Ignorance as a Self-Disciplining Device", *Review of Economic Studies*, 67, 529-544.

26. Carter, C., Braver, T., Barch, D., Botvinick, M. Noll, D. and J. Cohen (1998), "Anterior Cingulate Cortex, Error Detection, and the Online Monitoring of Performance", *Science*, 280, 747-749.

27. Damasio, A. (1994), *Descartes' Error: Emotion, Reason and the Human Brain*, New York: G.P. Putnam.

28. Elster, J. (2004), "Costs and Constraints in the Economy of the Mind", in I. Brocas and J.D. Carrillo eds. *The Psychology of Economic Decisions. Vol.2: Reasons and Choices*, pp. 3-14, Oxford: Oxford University Press.

29. Festinger, L. (1957), *A Theory of Cognitive Dissonance*, Stanford: Stanford U. Press.

30. Frederick, S., Loewenstein, G. and E. O'Donoghue (2002), "Time Discounting and Time Preference: A Critical Review", *Journal of Economic Literature*, 40(2), 351-401.

31. Fudenberg, D and D.K. Levine (2005), "A Dual Self Model of Impulse Control", forthcoming in *American Economic Review*.

32. Fudenberg, D. and J. Tirole (1991), *Game Theory*, Cambridge: MIT Press.

33. Gehring, W., Goss, B., Coles, M., Meyer, D. and E. Donchin (1993), "A Neural System for Error Detection and COmpensation", *Psychological Science*, 4(6), 385-390.

34. Gilbert, D. (2002), "Inferential Correction", in *Heuristics and Biases: the Psychology of Intuitive Judgment* T. Gilovich, D. Griffin and D. Kahneman, Eds., Cambridge, Cambridge University Press, 167-184.

35. Gul, F. and W. Pesendorfer (2001), "Temptation and Self-Control", *Econometrica*, 69, 1403-1435.

36. Gur, R. and H. Sackeim (1979), "Self-deception: A Concept in Search of a Phenomenon", *Journal of Personality and Social Psychology*, 37, 147-169.

37. Hall, R. and F. Mishkin (1982), "The Sensitivity of Consumption to Transitory Income: Estimates from Panel Data on Households", *Econometrica*, 50(2), 461-482.

38. Heath, C. and J. Soll (1996), "Mental Budgeting and Consumer Decisions", *Journal of Consumer Research*, 23(1), 40-52.

39. Kerns, J, Cohen, J., MacDonald, A., Cho, R., Stenger, A. and C. Carter (2004), "Anterior Cingulate Conflict Monitoring and Adjustments in Control", *Science*, 303, 1023-1026.

40. Knowlton, B., Mangels, J., and L. Squirre (1996), "A Neostriatal Habit Learning System in Humans", *Science*, 273, 1399-1402.

41. Laibson, D.I. (1997), "Golden Eggs and Hyperbolic Discounting", *Quarterly Journal of Economics*, 112, 443-477.

42. LeDoux, J. (1996) *The Emotional Brain. The Mysterious Underpinnings of Emotional Life*, Simon and Schuster: New York.

43. Loewenstein, G. (1996), "Out of Control: Visceral Influences on Behavior." *Organizational Behavior and Human Decision Processes*, 65, 272-292.

44. Loewenstein, G. and T. O'Donoghue (2005), "Animal Spirits: Affective and Deliberative Processes in Economic Behavior", *mimeo*, Carnegie Mellon and Cornell.

45. Loewenstein, G., Weber, W., Flory, J., Manuck, S. and M. Muldoon (2001), "Dimensions of Time-discounting", *mimeo*, Carnegie Mellon.

46. McClure, S., Laibson, D., Loewenstein, G. and J. Cohen (2004), "Separate Neural Systems Value Immediate and Delayed Monetary Rewards", *Science*, 306, 503-507.

47. Miller, E. and J. Cohen (2001), "An Integrative Theory of Prefrontal Cortex Function", *Annual Review of Neuroscience*, 24, 167-202.

48. Pascal, B. (1670), *Les Pensées*.

49. Poldrack, R. and P. Rodriguez (2004), "How Do Memory Systems Interact? Evidence from Human Classification Learning", *Neurobiology of Learning and Memory*, 82, 324-332.

50. Rabin, M. (1995), "Moral Preferences, Moral Constraints, and Self-Serving Biases", *mimeo*, UC Berkeley.

51. Rabin, M. (1998), "Psychology and Economics", *Journal of Economic Literature*, 36, 11-46.

52. Rauch, S., Whalen, P., Savage, C., Curran, T., Kendrick, A., Brown, H., Bush, G., Breiter, H. and B. Rosen (1997), "Striatal Recruitment during an Implicit Sequence Learning Task as Measured by Functional Magnetic Resonance Imaging", *Human Brain Mapping*, 5, 124-132.

53. Read, D., Loewenstein, G. and M. Rabin (1999), "Choice Bracketing", *Journal of Risk and Uncertainty*, 19(1), 171-197.

54. Robinson, T. and K. Berridge (2003), "Addiction", *Annual Review of Psychology*, 54, 25-53.

55. Shefrin, H.M and Thaler, R.H. (1988), "The Behavioral Life-Cycle Hypothesis", *Economic Inquiry*, 26, 609-643.

56. Simonson, I. (1990), "The Effect of Purchase Quantity and Timing on Variety Seeking Behaviour", *Journal of Marketing Research*, 32, 150-162.

57. Strotz, R.H. (1956), "Myopia and Inconsistency in Dynamic Utility Maximisation", *Review of Economic Studies*, 23, 166-180.

58. Thaler, R. (1985), "Mental Accounting and Consumer Choice", *Marketing Science*, 4, 199-214.

59. Thaler, R. (1990), "Anomalies. Saving, Fungibility, and Mental Accounts", *Journal of Economic Perspectives*, 4(1), 193-205.

60. Thaler, R.H., and H.M. Shefrin (1981), "An Economic Theory of Self-control", *Journal of Political Economy*, 89, 392-406.

61. Tirole, J. (2002), "Rational Irrationality: Some Economics of Self-Management", *European Economic Review*, 46, 633-655.

62. Whalen, P., Rauch, S., Etcoff, N., McInerney, S., Lee, M., and M. Jenike (1998), "Masked Presentations of Emotional Facial Expressions Modulate Amygdala Activity without Explicit Knowledge", *The Journal of Neuroscience*, 18(1), 411-418.

63. Zink, C., Pagnoni, G., Martin-Skurski, M., Chppelow, J., and G. Berns (2004), "Human Striatal Responses to Monetary Reward Depend on Saliency", *Neuron*, 42, 509-517.

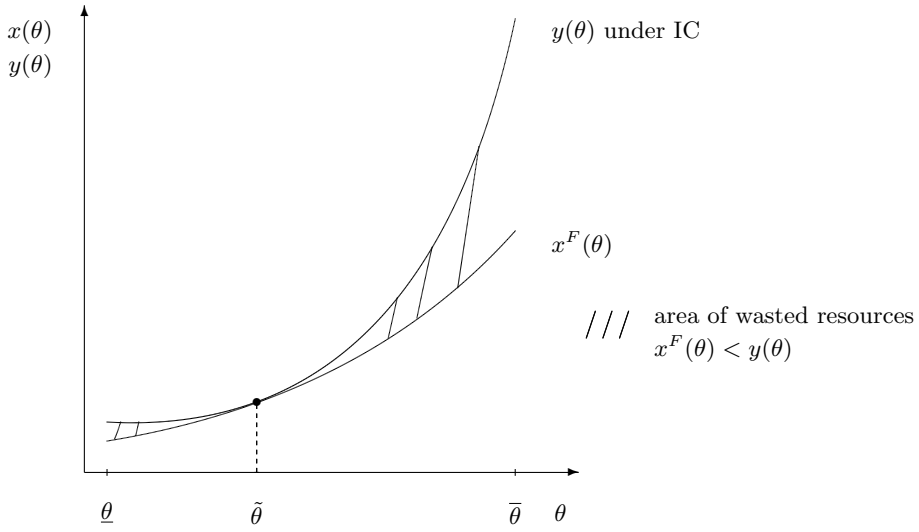

FIGURE 1A. OPTIMAL INCENTIVE COMPATIBLE CONTRACT

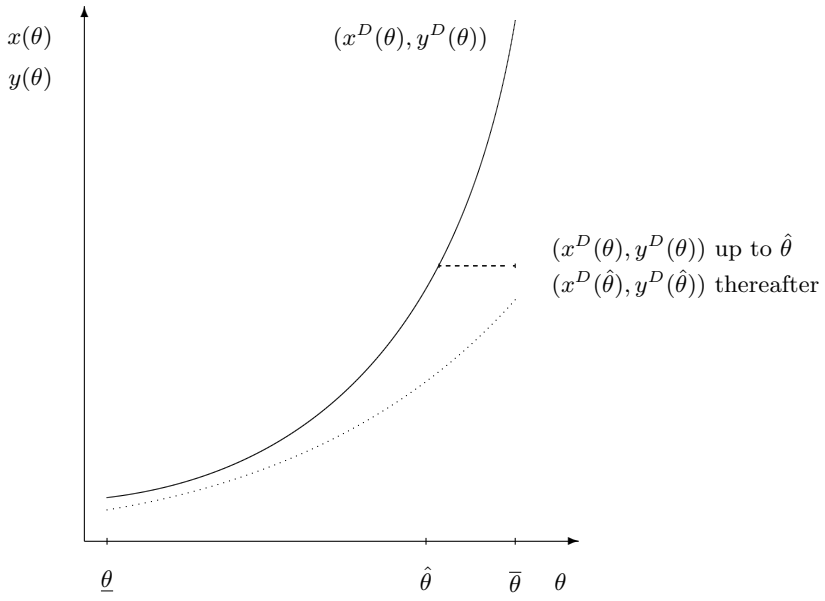

FIGURE 1B. FULL DELEGATION WITH AND WITHOUT CAP

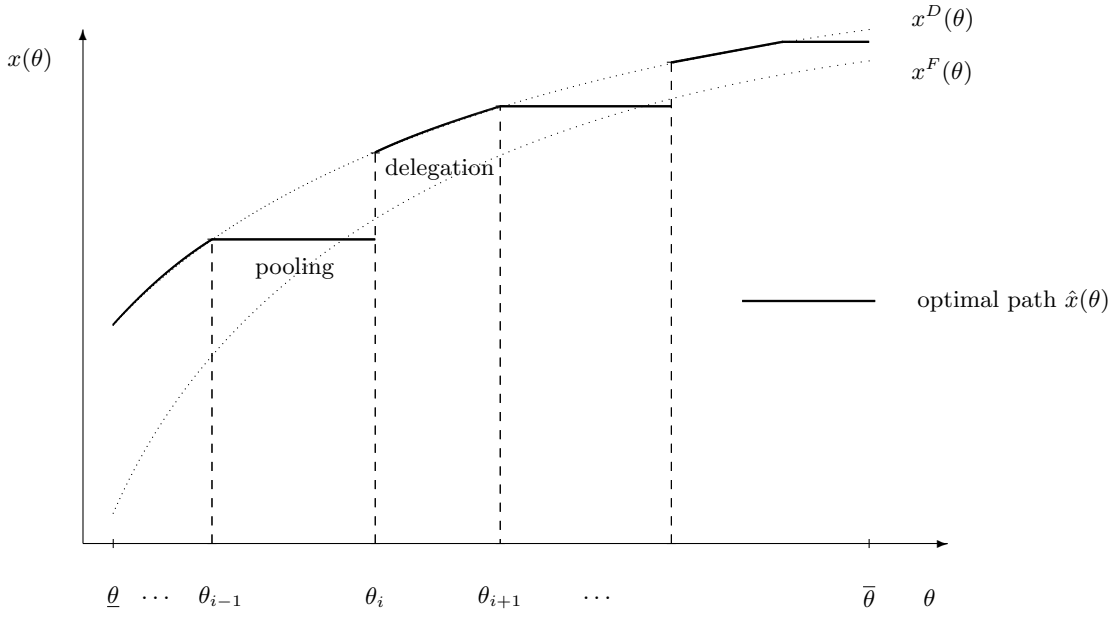

FIGURE 2. CONSUMPTION OF TEMPTING GOOD WHEN $T'' > 0$

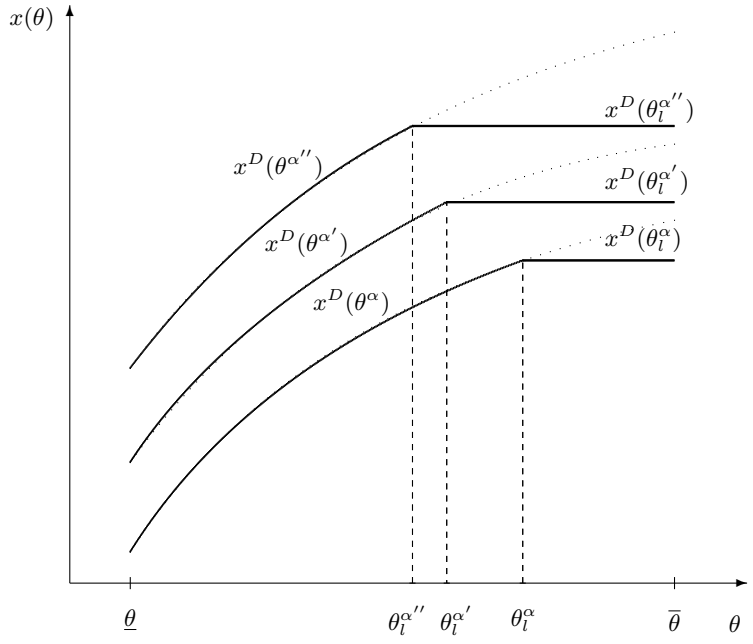

FIGURE 3. CONSUMPTION OF TEMPTING GOOD WITH $\alpha'' > \alpha' > \alpha > 1$