Optimal Taxation of Time Inconsistent Agents *

Pei Cheng Yu †

April 30, 2015

Abstract

This paper studies a model of taxation where the government faces time-inconsistent agents with private information on productivity. I study both non-sophisticated and sophisticated agents. Non-sophisticated agents are not fully aware of impending changes in their preferences, in contrast to sophisticated agents who can perfectly forecast their future preference changes. In particular, this paper addresses the problem of inadequate savings resulting from time-inconsistency. I demonstrate how the government can design an optimal truth-telling mechanism to achieve a full insurance outcome despite information asymmetry. In other words, when agents are time-inconsistent, the private information on productivity does not impede the government from implementing the first best allocation regardless of the sophistication level of the agents. For fully naïve agents, this is accomplished by deceiving them into telling the truth. For sophisticated agents, the government provides them with a commitment device in exchange for truth-telling. While for agents who are partially naïve, both types of mechanisms work. This result holds under very general conditions with agents who are time-inconsistent. I present a possible way to implement the optimal allocations using income specific non-linear savings taxes. As extensions, I examine economies with diversely naïve agents and when time consistent agents are present.

Keywords: Optimal taxation, time-inconsistency, Screening, Noncommon priors

JEL Classification Numbers: D03, D62, D82, D84, D86, D91, H21

*I thank David Rahman, Aldo Rustichini, Jan Werner, Manuel Amador, Martin Szydlowski, Kim-Sau Chung and Radek Paluszynski for many helpful comments. The views expressed herein are those of the authors and does not necessarily reflect those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

†University of Minnesota and FRB Minneapolis (e-mail: yuxxx680@umn.edu)
1 Introduction

Empirical evidence has shown that the behavioral biases exhibited in the real world are non-negligible and pervasive. Skeptics would argue that such behavioral bias may go away once we consider more important issues such as retirement savings or investment portfolios. However, evidence has shown the contrary. Evidence of behavioral anomalies suggests that there could be room for implementing paternalistic mechanisms to correct for time-inconsistent behavior. However, would implementing such corrective policies cause inefficiencies in other aspects of an individual’s economic behavior? The main interest of this paper is to examine how such policies interact with the incentives of an individual to work.

The optimal design of labor taxes to minimize its disincentive effects on labor supply is a key issue in public finance. Mirrlees (1971) introduced such a model with asymmetric information on production efficiency in workers, and derived a set of tax policies that could provide insurance for agents with the least distortion in effort provision. In addition to the efficiency and equity trade-off in a traditional Mirrlees environment, this paper introduces another concern when considering the appropriate labor tax: agents have time-inconsistent preferences with possibly limited awareness the preference reversal. Agents who are fully aware of their time-inconsistency are sophisticated, while those who are fully unaware are naïve. Agents who are aware of their time-inconsistency but are wrong about the severity or degree are called partially naïve. I study agents of all cognitive types. When agents are partially naïve of their time-inconsistency, a government wishing to help the agent ameliorate this behavioral bias (for example, to help the agent save enough for retirement) may inadvertently affect the agents' incentives to work. The main contribution of this paper is to describe the interaction between the adverse selection problem and the time-inconsistency problem, and characterize the optimal policy.

1 DellaVigna and Malmendier (2006) studied gym membership data and showed that 80% of monthly gym members would have been better off had they chosen to pay per visit. Ausubel (1999) and Shui and Ausubel (2005) have found similar biases in the credit card market. Gottlieb and Smetters (2013) have found similar evidence in the life insurance market. DellaVigna (2009) provides an overview of the empirical evidence for behavioral economics.

2 Benartzi and Thaler (2001) find that individuals do poorly when it comes to investment diversification. Madrian and Shea (2001) studied participation in the 401(k) plan and found evidence of strong default effects on the participation decision of individuals. O’Donoghue and Rabin (2001) show that a model of time-inconsistent individuals with naïveté can help explain the strong influence of the default option on retirement savings.

3 Sunstein and Thaler (2003) and O’Donoghue and Rabin (2003), (2006) have called for the government to implement policies that could help individuals make the right choice. The literature usually takes a cautious view of paternalism, even when supporting it, see Sunstein and Thaler (2008) and Camerer et al. (2003).
I find that the optimal set of allocations differs from the Mirrlees allocations. In particular, the government can avoid distorting the labor provision of any agent of any cognitive type and implement full insurance for skill realizations! More specifically, the main result shows that the government can achieve the first best allocation despite the presence of information asymmetry.

This surprising result is due to the fact that with naïve agents, the government can induce efficient labor provision by promising a large payoff in a certain good. However, after the preferences change, the agents no longer value the good they were promised, and would instead prefer the proportion of goods that corresponds to the first best allocation. In essence, the government can fool the naïve agents and does not need to deliver on the promise. As a result, the government can proceed to implement the full insurance policy without paying any information rent. With a well chosen fooling mechanism, the incentive to work is not hampered by the private information of the naïve agents.

To overcome information asymmetry with sophisticated agents, the government exploits the fact that sophisticated agents know about their time-inconsistency and have a demand for commitment devices. Using ideas developed in Chung (2015), the government will supply such a commitment device only if the agents report truthfully. If not, a misreporting agent will be tempted to choose an allocation designed to undo this commitment after the preferences change, and sophisticated agents can foresee this so they report truthfully. In other words, first best allocations are supported using credible off-equilibrium path threats. Since time-inconsistent agents will have a demand for commitment devices, as long as they are not fully naïve, screening with credible threats also works for both types of partially naïve agents.

As an extension, I consider an environment where both sophisticated and naïve agents coexist. In this setting, the government needs to screen the agents’ production efficiency and also their cognitive abilities. I show that mechanisms designed to fool more naïve agents and threaten more sophisticated ones can be combined without any efficiency loss. Therefore, the multidimensional screening problem does not alter the main result.

As an application, in light of recent empirical and experimental evidence, I focus on the problem of inadequate savings when agents have present bias. For implementation, the optimal fooling mechanism calls for the government to adopt an income-specific non-linear savings subsidy. An appropriate income specific regulation helps the government screen the productivity of the agents. The non-linearity of the savings policy helps fool or threaten the agents. For example, if the agents are non-sophisticated, before the present bias, they focus on a particular savings rate in the policy. However, when temptation for immediate gratification appears, the agents focus on a different savings rate in the policy.
The main results of this paper rely on the fact that the government is allowed to exploit the agents’ time-inconsistency. I consider an extension where the government is uncertain about the probability of a taste change, in essence, some agents could be time consistent. This limits the power of the government to fool or threaten the agents. I find that the presence of time consistent agents limits the main result. In particular, it is only optimal to fool the high productivity types, but not optimal for the low productivity types. I also find that social welfare increases as the population of time-inconsistent agents increases.

The benchmark model shows that these results apply to a very general setting with agents who experience preference changes and may or may not be accurate in their predictions of these changes. This demonstrates that the results may be applied more generally to different environments with dynamic inconsistencies and private information.

1.1 Related Literature

This paper is closely related to two strands of literature: the optimal taxation literature and the behavioral contracting literature. This paper aims to combine these two fields. Several works have already attempted to analyze the optimal government policies for maximizing the welfare of agents who suffer from temptation and self-control problems. Contrary to exploitative contracting, behavioral public economics aims to find the optimal paternal policy. For example, O’Donoghue and Rabin (2006) and Gruber and Koszegi (2004) have examined the utilization of government policies to curb addictive behavior. O’Donoghue and Rabin (2003) suggested using a mechanism design approach to find the most efficient policies when agents suffer from bounded rationality. This paper adopts such an approach. Several papers have also used a similar mechanism design or Ramsey type approach to characterize the optimal paternalistic government policies. A brief introduction of the papers most related to my work is given below.

This paper is closely related to Krusell, Kuruscu and Smith (2010) in that their work also studies the optimal taxation of consumers who suffer from temptation. They find that the government should subsidize future consumption in an effort to correct the agent’s impatience and tendency to save too little. This is in contrast to the no capital taxation result of Chamley-Judd, and is different from the usual no capital taxation in the mean presented in Golosov, Kocherlakota and Tsyvinski (2003) and Kocherlakota (2005). My work differs from Krusell, Kuruscu and Smith (2010) in two aspects. Firstly, I also consider non-sophisticated agents, while theirs are sophisticated. Secondly, their environment is a complete information one, while I introduce asymmetric information in productive efficiency.

4 Though this paper adopts a normative framework, it is not meant to support the implementation of paternalistic policies. A discussion of paternalism is provided in Section 8.
I introduce non-sophisticated agents because experimental evidence have found that humans are not fully aware of their own future preferences, and are also not very adept in learning about them. A discussion of naïveté is given in Section 8.

Amador, Werning and Angeletos (2006) have also examined government policies that could help agents with temptation. They study agents who suffer from temptation and are subject to future taste shocks. The government would like to help the agent overcome his temptation problems, but also allow the agent some room to accommodate the stochastic taste shock. However, the tendency to save too little confounds with the unobserved taste shock which creates a trade-off between commitment and flexibility. They find that a minimum savings rule is optimal. Their work also considers a sophisticated agent, and the adverse selection problem is in the agent’s taste shock. The main difference lies in the fact that my work seeks to explore how government policies aimed at helping a boundedly rational agent could distort his labor provision. Therefore, I have a production economy, while Amador, Werning and Angeletos (2006) focus on an endowment economy.

A few papers have studied the optimal taxation problem with asymmetric information and quasi-hyperbolic discounting agents. Bassi (2010) considers an environment where the hyperbolic discount factor is also non-observable, which creates a two-dimensional screening problem for the government. Guo and Krause (2015) study an environment where the government does not have full commitment. These two papers share a common goal with this one, but they all consider sophisticated agents or in settings where naïveté plays no role. This paper is the first to address the impact of cognitive limitations on tax design.

Several papers have examined taxation models where individuals are differentiated along two or more dimensions. Most closely related to this paper in terms of the policy issue it is concerned with is Diamond and Spinnewijn (2011). Their paper discusses a model with heterogeneity in both productivity and time preference. This paper is also concerned with the policy on savings, but the heterogeneity lies in the agents’ awareness of their underlying present bias. This type of multidimensional screening, where both the skill and cognitive ability of individuals are not observable, has not yet been analyzed in public policy or the economic literature.

The paper is organized as follows. Section 2 outlines the setup of the model. Section 3 and Section 4 and work out the results for our benchmark model for non-sophisticated and sophisticated agents respectively. Section 5 examines the optimal taxation of diversely naïve

---

More notably, Cremer, Pestieau and Rochet (2001) examine a model where both the productivity level and endowments are not observed by the government. Cremer, Pestieau and Rochet (2003) extend the model to an overlapping generations setting and endogenize individual endowments as inherited wealth. Beaudry, Blackorby and Szalay (2009) examines an economy where agents could participate in both market and non-market production and have different unobservable productivity levels for both sectors.
agents. Section 6 applies the results to examine the problem of inadequate savings. Section 7 explores a model where government is not certain about the probability of the event of a taste change. Section 8 discusses some of the impediments to the implementation of the proposed mechanism and other concerns. Section 9 concludes the paper. All proofs can be found in Appendix A.

2 The General Model

Following Spiegler [2011], I analyze the optimal allocation under two types of partial naïveté: magnitude naïveté and frequency naïveté. The benchmark model in this section will outline a general form of dynamic inconsistency. It can then be readily applied to a savings problem with present biased agents.

2.1 Setup of General Model

Consider an economy with $|N| \geq 2$ goods produced with labor or other goods and a continuum of agents denoted by the set $I = [0, 1]$. There are $|M| \geq 2$ types of agents denoted by the set $\Theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$. The types are distributed according to $\Pr(\theta = \theta_m) = \pi_m > 0$, for all $\theta_m \in \Theta$ with $\sum_{m=1}^{M} \pi_m = 1$.

The production of good $n$ depends on the labor input $l_n$ and the vector amount of other inputs $x_n \in \mathbb{R}_+^N$. Let $y_n = F_n(x_n, l_n; \theta_m) \in \mathbb{R}_+$ denote a continuous and differentiable production process strictly increasing in $l_n$ and $x_n$ of a type $m$ agent for good $n$. Let $l_n = G_n(y_n, x_n; \theta_j)$ denote the inverse of $F_n(x_n, l_n; \theta_j)$ with fixed input $x_n$. Each type of agent differs in their labor production efficiency: $F_n(x_n, l_n; \theta_j) > F_n(x_n, l_n; \theta_k)$, for any labor input $l_n > 0$ and $x_n$ and good $n$ with $\theta_j > \theta_k$. Therefore, a higher value of $\theta$ is associated with higher production efficiency. If a good does not depend on labor input, then it does not depend on $\theta$. The production of at least one of the goods requires labor.

As is standard in Mirrlees taxation, I assume that both the production efficiency of each agent and their labor input $l = (l_1, l_2, \ldots, l_N)$ are not observable by the government. The government can only observe output $y = (y_1, y_2, \ldots, y_N)$ and input $x$.

2.1.1 Consumer Utility

The agents have the following utility before consumption

$$U(c, l),$$
where $c = (c_1, c_2, \ldots, c_N)$. We will refer to $U$ as the ex-ante utility. This is the utility the agents use to evaluate their consumption plans. The agents’ utility changes when they are consuming to

$$V(c, l).$$

We will refer to $V$ as the ex-post utility. This utility models the tendency of a time-inconsistent agent to deviate from plans when confronted with an actual decision. Let $U$ and $V$ be continuously differentiable and let them be strictly increasing and concave in consumption: $\frac{\partial U}{\partial c_n} > 0$, $\frac{\partial^2 U}{\partial c_n^2} < 0$ and $\frac{\partial V}{\partial c_n} > 0$, $\frac{\partial^2 V}{\partial c_n^2} < 0$. Let $U$ be strictly decreasing and convex in labor: $\frac{\partial U}{\partial l_n} < 0$, $\frac{\partial^2 U}{\partial l_n^2} < 0$. Finally, for any good $n \in N$, let $\lim_{c_n \to 0} \frac{\partial U}{\partial c_n} = +\infty$ and $\lim_{c_n \to 0} \frac{\partial V}{\partial c_n} = +\infty$ to ensure an interior solution for consumption. Also, for any good $n \in N$, let $\lim_{l_n \to 0} \frac{\partial U}{\partial l_n} = 0$ and $\lim_{l_n \to +\infty} \frac{\partial U}{\partial l_n} = +\infty$ so the labor supply is always strictly positive and finite. Finally, I assume separability of consumption and leisure, so that the labor decision does not affect the marginal utility or marginal rates of substitution in goods consumption.

I will assume that the utility from consumption is different. More precisely, I assume the marginal rate of substitution for some consumption goods is different for $U$ than for $V$.

**Assumption 1** There exist $j, k \in N$ such that $\frac{\partial U}{\partial c_k} / \frac{\partial U}{\partial c_j} \neq \frac{\partial V}{\partial c_k} / \frac{\partial V}{\partial c_j}$.

First notice that since the utility is separable in consumption and labor, Assumption 1 is independent of the agents’ labor choice. Assumption 1 along with strictly increasing and concave utility implies a single crossing condition on the indifference curves for the ex-ante and ex-post utility of the two goods, $j$ and $k$. It allows the government to implement policies that would seem attractive to the ex-ante agent while remaining undesirable for the ex-post agent, or vice versa. If the ex-post and ex-ante utility satisfy Assumption 1, then the preference exhibits taste change. I will impose an additional standard assumption on the agents’ preferences: the marginal rate of substitution between consumption and output is smaller for more efficient agents.

**Assumption 2** For any good $n \in N$ that depends on labor for production, the ex-ante preferences satisfy the single crossing property: $\frac{\partial}{\partial \theta} \left( -\frac{\partial U}{\partial c_n} / \frac{\partial U}{\partial c_j} \right) < 0$ and $\frac{\partial}{\partial \theta} \left( -\frac{\partial V}{\partial c_n} / \frac{\partial V}{\partial c_j} \right) < 0$

### 2.1.2 Types of Non-sophistication

There are two common ways to model partial naïveté. Loewenstein, O’Donoghue and Rabin (2003) and Heidhues and Koszegi (2010) have interpreted partial naïveté as the underestimation of the magnitude of a taste change. Eliaz and Spiegler (2006) have interpreted partial naïveté as the underestimation of the likelihood of a taste change. Following
Spiegler (2011), I will refer to the former as magnitude naiveté and the latter as frequency naiveté.

**Definition 1** For some $\alpha \in (0, 1]$, agents are partially naïve in magnitude if, with probability one, they perceive their ex-post utility to be

$$W(c, l) = \alpha U(c, l) + (1 - \alpha) V(c, l).$$

Definition 1 defines magnitude naïveté. If $\alpha < 1$, then agents are certain that their preferences will change. However, since $\alpha$ is bounded away from 0, agents underestimate the degree of their taste change.

**Definition 2** Agents are partially naïve in frequency if they believe their ex-post utility to be $V(c, l)$ with probability $1 - \alpha$, where $\alpha \in (0, 1]$. In other words, let $W(c, l)$ denote the expected ex-post utility of the agent:

$$W(c, l) = \alpha U(c, l) + (1 - \alpha) V(c, l).$$

Definition 2 defines frequency naïveté. In essence, if $\alpha < 1$, the agents attach a positive probability to the likelihood of a change in the preference. However, since $\alpha$ is bounded away from 0, they underestimate the probability of their preferences changing.

Under both definitions, if $\alpha = 1$, the agents are fully naïve and never foresee the preference change. For both definitions, if $\alpha = 0$, the agents are sophisticated and are fully aware of their dynamic inconsistency. For both types of partial naïveté, I will refer to $\alpha$ as describing the sophistication level of an agent.

### 2.1.3 Timing

At date 0, the government designs the tax system. By the law of large numbers, the government knows the measure of each type of agent even before the agents learn their productivity types. At date 1, the agents’ types are realized. At date 2, the agents report their types according to reporting strategy $\sigma(\cdot)$. They make decisions according to the ex-ante utility from date 0 to date 2. At date 3, just when the agents are confronted with their consumption decisions, their preferences switch to the ex-post utility. At date 4, the agents make their consumption and labor decisions based on the ex-post utility. The timing of the model is shown in Figure [1].

There are three main stages in the timing which are highlighted by the actions of the government and agents. During the ‘design stage,’ the government designs the tax system. During
the ‘reporting stage,’ the agents choose the type they report to the government according to the ex-ante utility. During the ‘decision stage,’ the preferences of the agents switch and they choose consumption and labor supply according to the their reported type.

I assume that the government has full commitment; in essence, once the tax schedule is announced at date 0, the government is fully committed to carrying out the taxes as promised.

Using the revelation principle, the current timing focuses on a direct mechanism, where the agents report their productivity types and the government assigns allocations according to the reports. This method allows the analysis to focus on the constraints on achievable allocations and taxes coming solely from the informational frictions and time-inconsistency and free of ad hoc restrictions on tax instruments.

For tax implementation purposes, the timing can be dependent on the cognitive level of the agents. This will be discussed when implementation issues are presented.

2.1.4 Missing Markets

The lack of private insurance markets is a standard assumption in the Mirrlees taxation literature, because the presence of an efficient market that allows agents to insure against future skill shocks can make distortionary taxes redundant. I will assume that no private markets exist to insure against skill risks so the government has a role in insurance provision.

I will also assume that there are no markets for illiquid assets or other commitment devices. If such a market exists, then sophisticated agents can use it for commitment. The government would have a limited role in helping the sophisticated agents smooth consumption. However, for partially naïve agents, the presence of such markets does not preclude the need for government intervention. Heidhues and Koszegi (2009) demonstrate how commitment devices can do more harm than good for partially naïve agents, since they pay the cost of commitment but still suffer from self-control problems. In Section 8, I will discuss how the presence of a market for commitment can affect the results of this paper.
2.1.5 Welfare Criterion

The government has superior knowledge of the agents’ change in preference, and would try to help the agents commit to the ex-ante utility. Implicitly, I assume the non-sophisticated agents do not draw any inferences from the policies the government enacts. This is because they do not share the same prior as the government and are dogmatic in their beliefs.

The government evaluates allocations at date 0 according to the following welfare criteria

$$\sum_{m=1}^{M} \pi_m \psi [U(c(\theta_m), l(\theta_m))],$$

(1)

where \((c(\theta_m), l(\theta_m))\) denotes the allocation type \(m\) agent consumes. I assume that \(\psi \circ U\) is a strictly increasing and concave function, so that government has a desire to insure agents against productivity shocks. Notice that if \(\psi \circ U = U\), then the welfare criterion is utilitarian.

Much of the literature on dynamically inconsistent preferences have evaluated welfare with the ex-ante utility. I adopt the ex-ante utility relation as the main welfare criterion because it reflects the agents’ long-term planning, while the ex-post utility reflects the agents’ short-term temptations. In other words, the ex-post utility is not immune to regret and a benevolent government would consider the adverse implications if the agents give in to their urges. Under such perspectives, the choice of the welfare criteria is non-arbitrary, since the actions undertaken with respect to the ex-post preferences can be regarded as a systematic mistake the agents make, as in Bernheim and Rangel [2004]. I will show that the main idea of this paper is robust to changes in the welfare criteria in Section 8.

2.2 The Benchmark: No Private Information

In the benchmark, no private information case, the government maximizes social welfare subject to the government budget constraint

$$\sum_{m=1}^{M} \pi_m [F_n(x_n(\theta_m), l_n(\theta_m); \theta_m) - c_n(\theta_m)] = 0, \forall n \in N.$$

(2)

If the government is able to observe the productivity level of each agent, then it can achieve full insurance without distortions regardless of the agents’ time-inconsistency or their degree of naïvety. This is because with complete information, the agents work according to their skill type. The government can then choose an appropriate linear tax to correct the distortion caused by the taste change. The optimal allocation of the social planner problem without private information will be referred to as the first best allocation.
To illustrate how the government corrects the distortion caused by time-inconsistency, let 
\( c^*(\theta_m) = (c^*_1(\theta_m), c^*_2(\theta_m), \ldots, c^*_N(\theta_m)) \) denote the first best consumption for type \( \theta_m \). Also, let \( j, k \in N \) be any two goods such that \( \frac{\partial V}{\partial c^*_j} \neq \frac{\partial V}{\partial c^*_k} \). The first best consumption equates the marginal ex-ante utility across all goods,

\[
\frac{\partial U}{\partial c^*_{m,1}} = \frac{\partial U}{\partial c^*_{m,2}} = \ldots = \frac{\partial U}{\partial c^*_{m,N}}.
\]

Therefore, the government can choose linear taxes or subsidies \( \tau_j \) and \( \tau_k \) on goods \( j \) and \( k \) respectively such that

\[
\frac{1}{1 + \tau_j \frac{\partial V}{\partial c^*_{m,j}}} = \frac{1}{1 + \tau_k \frac{\partial V}{\partial c^*_{m,k}}},
\]

which implements the first best consumption allocation. For example, if the agents are tempted to consume too little of good \( j \) relative to good \( k \), then the government can subsidize good \( j \) or tax good \( k \).

There are three important features of the policy instrument in this environment. First, the taxes or subsidies can be linear. Second, notice that these linear instruments are the same across all productivity types. This is because all productivity types suffer from the same taste change. The linear taxes or subsidies also work independently of cognitive levels. This is because regardless of the sophistication level, the policy enacted to correct for the taste change does not distort the incentives to work. The government can always use a direct mechanism and ‘force’ the efficient agents to work more than the less efficient agents.

3 The Effects of Non-sophistication

Non-sophisticated agents hold erroneous beliefs about their future preferences, and hence make poor predictions about their behavior. This means that their reporting strategies do not reflect actual choices in the future, but a fictitious future-self. The government can proceed to exploit this mistaken belief of the non-sophisticated agents.

For this section, I will first describe the method the government uses to elicit truth-telling. I will then write down the optimization problem for both types of partial naïveté. The section ends with the main result which describes the optimal allocation for non-sophisticated agents.

3.1 Tax Instruments

The government is allowed to present a menu of tax options for the non-sophisticated agents. In the form of its resulting allocations, for a direct mechanism, the government issues
the following menu
\[
\left\{ \left( c^R(\theta_m), c^I(\theta_m) \right), \left( l^R(\theta_m), l^I(\theta_m) \right), \left( x^R(\theta_m), x^I(\theta_m) \right) \right\}_{\theta_m \in \Theta}.
\]

More concretely, after the tax schedule is announced, non-sophisticated agents of all types ‘mentally’ choose a set of allocations \((c^I(\theta), l^I(\theta), x^I(\theta))\). However, the non-sophisticated agent will ‘actually’ choose allocation \((c^R(\theta), l^R(\theta), x^R(\theta))\) to maximize the ex-post utility. The superscript \(I\) represents ‘imaginary’, since it is never actually chosen, but were the perceived choices before the preference switch. The superscript \(R\) represents ‘reality,’ since they are actually chosen after the preference change, but were not planned when the agents made their reports. The government will attempt to exploit the unforeseen taste change. I will set \(l^R(\theta_m) = l^I(\theta_m) = l(\theta_m)\) and \(x^R(\theta_m) = x^I(\theta_m) = x(\theta_m)\), and show that \(c^R(\theta) \neq c^I(\theta)\) is enough to exploit non-sophistication.

For both types of partial naïveté, agents do not fully anticipate their preference change before the consumption stage, which makes the design of both ‘imaginary’ and ‘real’ allocations matter. However, the imaginary allocations are evaluated differently under the two types of partial naïveté. For magnitude naïveté, agents make their reporting decision based on \(U(c, l)\) while anticipating a taste change of \(W(c, l)\). Therefore, they require \(c^I(\theta)\) to be more appealing than \(c^R(\theta)\) under \(W(c, l)\), and the reporting strategy is evaluated using \(U(c, l)\). While for frequency naïveté, agents make their reporting decision based on their expected ex-post utility \(W(c, l)\). More specifically, they require \(c^I(\theta)\) to be more appealing than \(c^R(\theta)\) under \(U(c, l)\), and the reporting strategy is evaluated at the expected utility of \(U(c, l)\). Notice that for both types of partial naïveté, the real allocation is more appealing than the imaginary allocation under the ex-post utility.

### 3.2 The Planning Problem

With full commitment by the government, the revelation principle implies the government can focus on a direct mechanism that elicits truth telling. With a slight abuse of notation, I will define \(G(y(\theta_{m'}), x(\theta_{m'}); \theta_m) \in \mathbb{R}^N\) as the labor input vector for a type \(\theta_m\) agent with reporting strategy \(\sigma(\theta_m) = \theta_{m'}\).

The two types of partial naïveté represent two different perspectives on how a cognitively limited agent evaluates the future preference, which affects their reporting strategy. As a result, incentive compatibility is different under the different types of partial naïveté.
3.2.1 Planning Problem with Magnitude Naïveté

Under magnitude naïveté, the government’s problem is to choose an allocation menu \( \{c^R(\theta_m), c^I(\theta_m), l(\theta_m), x(\theta_m)\}_{m=1}^M \) to maximize (1) subject to the government budget constraint (2) evaluated at the real allocations and

\[
U(c^I(\theta_m), l(\theta_m)) \geq U(c^I(\theta_m), G(y(\theta_m'), x(\theta_m'); \theta_m)), \forall \theta_m, \theta_m' \in \Theta, \theta_m \neq \theta_m',
\]

(3)

\[
W(c^I(\theta_m), l(\theta_m)) \geq W(c^R(\theta_m), l(\theta_m)), \forall \theta_m \in \Theta;
\]

(4)

\[
V(c^R(\theta_m), l(\theta_m)) \geq V(c^I(\theta_m), l(\theta_m)), \forall \theta_m \in \Theta.
\]

(5)

Constraint (3) is the incentive compatibility constraint. Notice that the agents’ reporting strategy is determined by \( W \), which is evaluated at the imaginary allocation. This is because constraint (4) makes sure the agents perceive they would choose the imaginary allocation for any effort level, even if they deviated from truth-telling. Constraints (4) and (5) are referred to as the fooling constraints, which ensure the agents focus on the imaginary allocations when deciding their reporting strategy and choose the real allocations after their taste change.

In the magnitude naïveté interpretation, the agents are certain that their preference will change, and they anticipate that change by evaluating their reporting strategy using the imaginary allocation. This is because they believe that once their preferences switch to \( W(c, l) \), they would prefer the imaginary allocation over the real allocation. Fooling constraint (4) ensures this. However, the agents underestimate the magnitude of the taste change and, if fooling constraint (5) is satisfied, would instead prefer to choose the real allocation over the imaginary allocation after the preference switch.

3.2.2 Planning Problem with Frequency Naïveté

For frequency naïveté, the government chooses \( \{c^R(\theta_m), c^I(\theta_m), l(\theta_m), x(\theta_m)\}_{m=1}^M \) to maximize (1) subject to the government budget constraint (2) evaluated at the real allocations and the fooling constraint (5) with the following incentive compatibility constraint

\[
\alpha U(c^I(\theta_m), l(\theta_m)) + (1 - \alpha) U(c^R(\theta_m), l(\theta_m)) \geq \alpha U(c^I(\theta_m'), G(y(\theta_m'), x(\theta_m'); \theta_m)) + (1 - \alpha) U(c^R(\theta_m'), G(y(\theta_m'), x(\theta_m'); \theta_m)),
\]

(6)

and the fooling constraint for the imaginary allocation

\[
U(c^I(\theta_m), l(\theta_m)) \geq U(c^R(\theta_m), l(\theta_m)), \forall \theta_m \in \Theta.
\]

(7)
The difference between frequency naïveté and magnitude naïveté lies in the beliefs of the future preference. In frequency naïveté, the agents believe with some probability \( \alpha \) that they will choose the imaginary allocation evaluated at the ex-ante preference \( U(c, l) \), which is represented in \( (7) \). This is different from the fooling constraint \( (4) \) for magnitude naïveté, where the agents are certain that their preferences would change, but underestimate the extent of this shift.

### 3.2.3 More on the Constraints

If there exists a type \( m \) such that at least one of the fooling constraints is non-binding, then \( c^I(\theta_m) \neq c^R(\theta_m) \), and the government is fooling the type \( m \) agent. In other words, the government is exploiting the agents’ non-sophistication by using a fooling mechanism.

Notice the imaginary allocations are not required to satisfy the government budget constraint. This is because the government only cares about the real allocation, and views the imaginary allocations as an empty promise. The government is certain about the degree of the naïveté and present bias of the agents, so it places no weight on a future where it may need to actually honor the delivery of imaginary allocation. Another concern is that the agents do not realize the aggregate imaginary allocation violates the government budget constraint. This is because each agent is infinitesimally small, and even though an agent believes he would consume the imaginary allocation, he does not consider what other agents believe and how they would behave.

### 3.3 Main Result for Non-sophisticated Agents

By Assumption 1 and Assumption 2, it can be shown that the government can achieve the first best allocation. In other words, surprisingly, private information does not matter in an environment where all agents harbor some naïveté.

**Proposition 1**  
*The optimal allocation for the environment where agents have private information on productivity and are fully naïve or partially naïve in magnitude or frequency about their preference changes is the same as the allocation in the environment without private information.*

Proposition 1 states that the private information problem can be alleviated if the agents are not sophisticated. This is because the government can enact policies to fool the agents into believing a particular allocation would be realized in the future, which can provide the necessary incentives for the agents to report truthfully. After their preferences change, the duped agents would find the first best allocation superior to the imaginary allocation.
In other words, with the imaginary allocations, the government is able to provide the information rents necessary for truth-telling. However, these rents are imaginary. After the preference of the agents change, the government is able to implement the first best allocation without paying the information rents. Indeed, it necessarily follows that it is optimal for the government to deceive the agents when they are not sophisticated.

**Corollary 1** If \( \alpha > 0 \), it is optimal for the government to implement a fooling mechanism.

The key to deceiving the agents is to load the rents on goods that they value during the reporting stage, but would not value as much relative to other goods after the preference change. By Assumption 1, suppose \( \frac{\partial U}{\partial c_k} / \frac{\partial U}{\partial c_j} > \frac{\partial V}{\partial c_k} / \frac{\partial V}{\partial c_j} \), then the agents value good \( k \) more than good \( j \) at the reporting stage. The government can then promise more of good \( k \) than good \( j \) for the imaginary allocations to elicit truthful reports as long as the agents hold the wrong beliefs. However, after the preference change, the promise of more good \( k \) is less appealing, and the agents would no longer choose the imaginary allocations but the real allocations, with less of good \( k \).

It is interesting to note that there is a discontinuity in the optimal welfare with respect to the cognitive limitations of the agents. With magnitude naïveté, the government is able to achieve first best welfare for any sophistication level \( \alpha \in (0, 1] \). However, with fully sophisticated agents \( (\alpha = 0) \), the fooling mechanism can only implement the Mirrlees allocations which requires information rent for the productive types. This is because the sophisticated agents would foresee perfectly their taste change at the consumption stage and would thus be immune to any deceptions at the reporting stage. As a result, there is a discontinuity in welfare which is similar to the discontinuity in Heidhues and Koszegi (2010).

A more surprising result for partial naïveté is that this discontinuity is present for naïveté in frequency as well. Spiegler (2011) has shown the optimal contract to be continuous with respect to cognitive limitations in a second-degree price discrimination setting for frequency naïveté. However, Proposition 1 shows that this continuity result does not hold in the optimal taxation setting even with naïveté in frequency.

However, the discontinuity is only present if the government uses a fooling mechanism. For sophisticated agents, the government can also exploit their time-inconsistency and attain full insurance without any distortions. In the next section, I will present such a mechanism.

### 4 The Effects of Sophistication

The fooling mechanism introduced for non-sophisticated agents no longer work for sophisticated agents. Sophisticated agents are immune to deception, and will require actual
information rents for full revelation of their types under a fooling mechanism.

However, the government can design an off-equilibrium threat that will be chosen only if an agent misreports. This threat helps attain first best allocations. The idea is that the government can double check an agent’s reported type with the ex-post preferences. Sophisticated agents respond to threats because they want a commitment device, and the government is willing to provide it as long as the agents are truthful. Adverse selection models with sophisticated agents have been explored in Esteban and Miyagawa [2005] and Chung [2015]. In particular, I rely on insights provided in Chung [2015].

4.1 Tax Instruments

In the form of allocations, the government introduces the following menu for sophisticated agents

\[ \{ (c^R(\theta_m), c^T(\theta_m)), (l^R(\theta_m), l^T(\theta_m)), (x^R(\theta_m), x^T(\theta_m)) \}_{\theta_m \in \Theta}. \]

When the sophisticated agents are choosing their reporting strategy, they know that their preferences will change and they will be tempted to consume a set of allocations that is deemed undesirable as measured by their ex-ante utility \( U \). The government takes advantage of the agents’ time-inconsistency and their awareness by introducing a set of allocations \((c^T(\theta), l^T(\theta), x^T(\theta))\) such that, after the preference change, an agent that reports truthfully would never choose it, but an agent who misreports would. The superscript \( T \) represents ‘threat,’ because the misreporting agent would consider this allocation to be inferior according to the ex-ante preferences. I will refer to this allocation as the threat allocation. Adding the threat allocation to the menu deters the agents from misreporting.

It is important to note that threats are non-credible if \( l^R(\theta_m) = l^T(\theta_m) \) and \( x^R(\theta_m) = x^T(\theta_m) \). This is because all agents share the same preference for goods consumption and a misreporting agent would not be caught if the production plans are the same for the real and threat allocations. As a result, the real allocations and threat allocations have to be different in both consumption and production.

4.2 The Planning Problem

The planning problem of the government maximizes (1) subject to the government budget constraint (2) evaluated at the real allocations and the following incentive compatibility constraint, \( \forall \theta_m, \theta_{m'} \in \Theta, \theta_m \neq \theta_{m'} \),

\[ U(c^R(\theta_m), l^R(\theta_m)) \geq U \left[ c^T(\theta_{m'}), G(y^T(\theta_{m'}), x^T(\theta_{m'}); \theta_m) \right], \quad (8) \]
and the credible threat constraints \( \forall \theta_m, \theta_{m'} \in \Theta, \theta_m \neq \theta_{m'} \),

\[
V(c^R(\theta_m), l^R(\theta_m)) \geq V(c^T(\theta_m), l^T(\theta_m)),
\]

\[
V(c^T(\theta_{m'}), l^T(\theta_{m'})) \geq V(c^R(\theta_{m'}), l^R(\theta_{m'})),
\]

When preferences change, credible threat constraint (9) ensures the agents choose the real allocations provided that they tell the truth. By credible threat constraint (10), if the agents misreport their production capabilities, they will be tempted to choose the threat allocations. However, the incentive compatibility constraints make sure that the real allocations are weakly preferred to the threat allocations according to the ex-ante utility.

If there exists a type \( \theta_m \) agent who has at least one of the credible threat constraints slack, then \( (c^R(\theta_m), l^R(\theta_m), x^R(\theta_m)) \neq (c^T(\theta_m), l^T(\theta_m), x^T(\theta_m)) \), and the government is threatening the type \( \theta_m \) agent through a threat mechanism. The threat allocations are not required to satisfy the government budget constraint, because sophisticated agents are capable of using backward induction and the threat allocations are on the off-equilibrium path.

### 4.3 Main Result for Sophisticated Agents

Using Assumption 1 and Assumption 2, the following proposition shows that private information does not matter in an environment with sophisticated agents as well!

**Proposition 2** The optimal allocation for the environment where agents have private information on productivity and are fully aware of their preference changes is the same as the allocation in the environment without private information.

Proposition 2 states that even with sophisticated agents, the private information does not cause a distortion to the government’s planning problem. This is because the sophisticated agents are aware that their future self would distort their consumption plan, and would desire a commitment device that deters them from doing so. The government provides this commitment device and asks the sophisticated agents to report truthfully in return. The government caters to the temptations of the ex-post preferences if the agents misreport. In essence, the government holds the agents’ future self hostage and threaten to distort the consumption plan unless the agent reports truthfully. The threat is credible because the agents’ future self has the same productivity efficiency.

Observe that no matter the sophistication level of the agents, as long as they are time-inconsistent, the government is able to achieve first best welfare. Discontinuity is only present if the mechanism is not malleable with the sophistication level of the agents. This raises a
subtle issue regarding the difference in mechanisms for varying degrees of naïveté. A fully naïve agent has to be fooled, since they do not respond to threats. While a sophisticated agent has to be threatened, because they can never be deceived. Since partially naïve agents have demand for commitment devices too, they are not immune to threats as well.

For magnitude naïveté, the threats are evaluated using the erroneous ex-post utility \( W \), so the credible threat constraints \( \forall \theta_m, \theta_m' \in \Theta, \theta_m \neq \theta_m' \) are

\[
W(c^R(\theta_m), l^R(\theta_m)) \geq W(c^T(\theta_m), l^T(\theta_m)),
W\left[c^T(\theta_m'), G(y^T(\theta_m'), x^T(\theta_m'); \theta_m)\right] \geq W\left[c^R(\theta_m'), G(y^R(\theta_m'), x^R(\theta_m'); \theta_m)\right].
\]

The government also needs to make sure that after the preferences change, the truthful agents will indeed choose the real allocation, so an additional constraint is needed:

\[
V(c^R(\theta_m), l^R(\theta_m)) \geq V(c^T(\theta_m), l^T(\theta_m)), \forall \theta_m \in \Theta.
\] (11)

Constraint (11) makes sure that the government’s threat does not back fire.

For frequency naïveté, the credible threat constraints remain the same. However, the incentive compatibility constraints \( \forall \theta_m, \theta_m' \in \Theta, \theta_m \neq \theta_m' \) are

\[
U\left(c^R(\theta_m), l^R(\theta_m)\right) \geq \alpha U\left[c^R(\theta_m'), G\left(y^R(\theta_m'), x^R(\theta_m'); \theta_m\right)\right] \\
\quad + (1 - \alpha) U\left[c^T(\theta_m'), G\left(y^T(\theta_m'), x^T(\theta_m'); \theta_m\right)\right].
\]

The threat allocation makes sure that the partially naïve agent does not think misreporting is worth the risk of consuming the threat allocation. The following corollary shows that partially naïve agents can be screened using a threat mechanism and Table 1 summarizes the application of the two types of mechanisms and their applicability to each cognitive type.

**Corollary 2** It is optimal to threaten the agents when \( \alpha < 1 \).

<table>
<thead>
<tr>
<th></th>
<th>Fully Naïve</th>
<th>Partially Naïve</th>
<th>Sophisticated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fooling</td>
<td>✓</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td>Threat</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 1: Summary of Mechanism for Different Cognitive Types
5 Model with Diversely Naïve Agents

The previous model had agents differing in their production efficiency while sharing the same cognitive features. In this section, I consider the setting where a government faces dynamically inconsistent agents who differ in their cognitive abilities, which is hidden from the government. Since both productivity and sophistication level are unobserved by the government, the optimal policy has to solve a multidimensional screening problem. The government would like to know which agents are productive so they could be encouraged to produce more. However, the form of the incentive scheme would depend on the sophistication level of each agent.

The sophistication level of agents is distributed within the bounded support of $[0, 1]$. Let $\Pi(\theta_m, \alpha)$ denote the joint distribution of productivity and sophistication level. It is easy to show that as long as all agents are bounded away from either full sophistication or full naïveté, the government is still capable of achieving the first best allocation. I will refer to mechanisms that attain the first best allocation as effective.

**Lemma 1** A fooling mechanism that is effective for agents of sophistication level $\alpha$ is also effective for more naïve agents. A threat mechanism that is effective for agents of sophistication level $\bar{\alpha}$ is also effective for more sophisticated agents.

Lemma 1 shows that a fooling mechanism for sophistication level $\alpha$ agent can also fool agents with $\alpha > \alpha$. While a threat mechanism for sophistication level $\bar{\alpha}$ agent, can also credibly threaten agents with $\alpha < \bar{\alpha}$.

A fooling mechanism for $\alpha$ will work for any $\alpha > \alpha$, regardless of the joint distribution of productivity and cognitive limitation. This is because providing incentives for the least naïve agents for truth-telling is the most difficult, so any incentives that could separate the productivity of the least naïve agents will also be truth-telling for more naïve agents. Similarly, a threat mechanism for $\bar{\alpha}$ will work for any $\alpha < \bar{\alpha}$, since the least sophisticated agent needs the strongest threat to be willing to divulge his true production capability, so the threat would also work for more sophisticated agents.

With Lemma 1, the government can choose an arbitrary target sophistication level, $\hat{\alpha}$, such that all agents who are more sophisticated than $\hat{\alpha}$ are threatened by using the same threat mechanism and those who are more naïve are fooled by using the same fooling mechanism. I will henceforth refer to the agents with sophistication level $\alpha \in (\hat{\alpha}, 1]$ as relatively naïve and agents with sophistication level $\alpha \in [0, \hat{\alpha})$ as relatively sophisticated. Hence, a fooling mechanism designed for agents with sophistication level $\hat{\alpha}$ is applied to relatively naïve agents and a threat mechanism designed for agents with sophistication level $\hat{\alpha}$ is ap-
plied to relatively sophisticated agents. Lemma 1 ensures separation of productivity types through bisecting the population of agents according to their cognitive limitations.

The only concern is whether implementing a fooling mechanism and threat mechanism simultaneously would impede separation in sophistication levels. In particular, a more sophisticated agent needs to be prevented from pretending to be a less sophisticated agent. The government can choose a fixed target sophistication level at \( \hat{\alpha} \in (0, 1) \) and introduce the following menu

\[
\{(c_R(\theta_m), c_I(\theta_m), c_T(\theta_m)), (l_R(\theta_m), l_I(\theta_m), l_T(\theta_m)), (x_R(\theta_m), x_I(\theta_m), x_T(\theta_m))\}_{\theta_m \in \Theta}.
\]

The imaginary and threat allocations are chosen such that agents with sophistication level \( \hat{\alpha} \) are fooled and threatened with effective mechanisms. I will refer to this mechanism as a hybrid mechanism.

By the fooling constraints and the credible threat constraints, a relatively naïve agent would evaluate the merits of truth-telling at the imaginary allocations, while the merits of misreporting would be evaluated at the imaginary or threat allocations. However, as long as the imaginary benefits from truth-telling are sufficiently appealing, a relatively naïve agent would divulge both productivity and sophistication levels truthfully. On the other hand, a relatively sophisticated agent evaluates the merits of truth-telling at the real allocations, while the merits of misreporting would be evaluated at the threat allocations. Therefore, with a sufficiently strong threat, a relatively sophisticated agent would divulge both productivity and sophistication levels truthfully.

**Proposition 3** The optimal allocation for the environment with diversely naïve agents where agents have private information on productivity and cognitive limitations are not observable is the same as the allocation in the environment without private information.

Proposition 3 follows immediately from the fact that relatively naïve and relatively sophisticated agents can be costlessly separated using a hybrid mechanism. This is because the relatively naïve agents focus on the imaginary allocations for truth-telling, while the relatively sophisticated agents focus on the threat allocations for misreporting. Therefore, the fooling and threat mechanisms do not interact when they are integrated, so the government can consider each cognitive population separately. Along with Lemma 1, the government can choose any arbitrary target sophistication level \( \hat{\alpha} \), so the population that is being fooled (or threatened) can be arbitrary.

Another feature of Proposition 3 is that it relies on very little information about the economic environment. The government does not need to know the joint distribution of
sophistication level and productivity level, which is an integral information in usual multi-
dimensional screening mechanisms.\(^6\)

6 The Savings Problem

I will now consider a special case of the general model. The agents live for two periods. They produce and make consumption and savings decision in the first period, and consume their savings in the second period.

Following the usual Mirrlees setup, the production technology is linear, \(F(l; \theta_j) = \theta_j l\). Therefore, in a competitive equilibrium, the wages are equated to the marginal productivity of labor. There is also a storage technology that transfers one unit of good in the first period to one unit of second period good. (Alternatively, they have access to a bond with interest rate 0.)

The agents have the following ex-ante utility

\[ U(c, k, l) = u(c) - h(l) + w(k). \]

They face the following ex-post utility function

\[ V(c, k, l) = u(c) - h(l) + \beta w(k). \]

The period utilities are continuously differentiable and satisfy the usual concavity assumptions \(u', -u'' > 0\) and \(w', -w'' > 0\), while the dis-utility from labor satisfies \(h', h'' > 0\). Also, \(\lim_{c \to 0} u'(c) = +\infty\) and \(\lim_{k \to 0} w'(k) = +\infty\), so consumption in both periods will be strictly positive.

I will focus on the case with present bias, where \(\beta < 1\). A smaller \(\beta\) represents a stronger bias for present consumption. I will refer to \(\beta\) as measuring the degree of temptation the agents suffer from. Following O’Donoghue and Rabin [2001], a partially naïve agent in magnitude perceives his degree of present bias to be \(\hat{\beta} \in (\beta, 1]\) before reporting his type. Notice if \(\hat{\beta} = 1\), then the agent is fully naïve and unaware of his present bias. If \(\hat{\beta} = \beta\), then the agents are sophisticated. Similar to the general model, the partially naïve agents’ perceived present bias is always strictly greater than the actual present bias, \(\hat{\beta} > \beta\), so the agents underestimate the degree of their present bias.

For frequency naïveté, a partially naïve agent believes that his preferences change to \(\beta\) with probability \(1 - \alpha\) and it would stay the same with probability \(\alpha\). If \(\alpha = 1\), then the

\(^6\) For an introduction to the multi-dimensional screening model, see Armstrong and Rochet [1999]
agent is fully naïve, and if $\alpha = 0$, then the agent is sophisticated. A partially naïve agent corresponds to a belief where $\alpha \in (0, 1)$ and underestimates the probability of the present bias occurring.

In the current setup, the present bias is similar to a temptation shock that triggers immediate gratification and under-weighs the virtues of saving. This formulation has an equivalent quasi-hyperbolic discounting representation under the timing shown in Figure [1].

The three period quasi-hyperbolic representation shown in Figure [1] is

$$U_1(c, k, l) = \hat{\beta} \delta [u(c) - h(l) + \delta w(k)],$$

$$U_2(c, k) = u(c) - h(l) + \hat{\beta} \delta w(k),$$

$$U_3(k) = w(k),$$

with $\delta = 1$. In the first period, the agents enroll in an income-specific savings plan (with several options available in the plan). In the second period, agents work to meet the income requirements of the plan and make consumption and savings decision according to the plan. Finally, in the third period, agents consume their retirement savings. If $\hat{\beta} = \beta$, then the agents are sophisticated and the transformed model is similar to Laibson [1997], and if not, then it is similar to the model with cognitive limitations as presented in O’Donoghue and Rabin [2001].

A full fledged quasi-hyperbolic discounting model is presented in Appendix B. For expository purposes, I will present the two-period model and discuss non-sophisticates and sophisticates separately.

6.1 Savings with Non-sophisticates

The savings problem is a simplified setup of the general model with preference changes. The government’s planning problem is also similar. Note that Assumption 1 and Assumption 2 are automatically satisfied.

**Corollary 3** It is optimal to use a fooling mechanism. The optimal allocation for the environment with private information and non-sophisticates is the same as the allocation in the environment without private information.

To demonstrate how Corollary 3 works, consider an economy with two productivity types $\Theta = \{\theta_b, \theta_g\}$, where $\theta_g > \theta_b$, and let the government be utilitarian, so that $\psi \circ U = U$. By Corollary 3, $(c^R, k^R, l)$ will be the first best allocation. The first best consumption and savings are equated across types with perfect consumption smoothing across periods. Also,
the marginal cost of effort is equated to the marginal benefit of consumption. As a result, the efficient agents work more than the inefficient agents. For simplicity, I will examine the fully naïve case, so there is no distinction between magnitude and frequency naïveté.

The incentive compatibility constraints are

\[
\begin{align*}
u(c_I^g) - h(l_g) + w(k_I^g) &\geq \nu(c_I^b) - h\left(\frac{\theta_bl_b}{\theta_g}\right) + w(k_I^b), \\
u(c_I^b) - h(l_b) + w(k_I^b) &\geq \nu(c_I^g) - h\left(\frac{\theta_gl_g}{\theta_b}\right) + w(k_I^g),
\end{align*}
\]

and the fooling constraints are

\[
\begin{align*}
u(c_g^R) + \beta w(k_g^R) &\geq \nu(c_g^I) + \beta w(k_g^I), \\
u(c_b^R) + \beta w(k_b^R) &\geq \nu(c_b^I) + \beta w(k_b^I), \\
u(c_g^I) + w(k_g^I) &\geq \nu(c_g^R) + w(k_g^R), \\
u(c_b^I) + w(k_b^I) &\geq \nu(c_b^R) + w(k_b^R).
\end{align*}
\]

In Figure 2, the flatter solid (blue) curve represents the indifference curve of the ex-ante utility at allocation \( (c^R, k^R) \). The steeper solid (red) curve represents the indifference curve of the ex-post utility at allocation \( (c^R, k^R) \). The imaginary allocations have to be in the area bounded by the solid line indifference curves in the north-west region. (If \( \beta > 1 \), then the

Figure 2: Finding the Imaginary Allocations
imaginary allocations would be bounded by the solid line indifference curves in the south-east region.) Any allocation within this area satisfies the inequalities \((14), (15), (16)\) and \((17)\). Furthermore, the incentive compatibility constraints, \((12)\) and \((13)\), provide an upper and lower bound to the difference in ex-ante utility of the two types of agents. In essence,

\[
h\left(\frac{\theta_m l_g}{\theta_b}\right) - h(l_b) \geq \left[u(c_g^l) + w(k_g^l)\right] - \left[u(c_b^l) + w(k_b^l)\right] \geq h(l_g) - h\left(\frac{\theta_b l_b}{\theta_g}\right).
\]

Therefore, given the first best labor provision, the imaginary allocations have to be within the dashed indifference curves, where the good type’s imaginary allocation is within the bold dashed area, and the bad type’s is within the light dotted area.

### 6.1.1 Implementation: Income Taxation and Retirement Savings

To implement the first best allocation in this environment, the government can rely on savings subsidies that are non-linear and productivity specific. Consider the fully naïve case with two productivity types and an utilitarian government. Denote the first best allocation as \((c^*, k^*, l^*_m)_{m \in \Theta}\), which are the real allocations the government wishes to implement. They satisfy the following marginal conditions: intertemporal substitution \(u'(c^*) = w'(k^*)\), full insurance across types \(c_g^* = c_b^* = c^*\) and \(k_g^* = k_b^* = k^*\), and intratemporal substitution \(u'(c^*) = \frac{1}{\theta_m} v'(l^*_m)\). It also satisfies the government budget constraint: \(\pi g \theta_g l_g^* + \pi b \theta_b l_b^* = c^* + k^*\).

Therefore, the real savings subsidies are the same for all types and are chosen to smooth consumption across periods optimally:

\[
1 + \tau^* = \beta.
\]

The lump-sum taxes have to satisfy the government budget constraint: for the high productivity agents,

\[
\Lambda_g^* = \pi_b (\theta_g l_g^* - \theta_b l_b^*) + (1 - \beta)k^*,
\]

and for the low productivity agents,

\[
\Lambda_b^* = -\pi_g (\theta_g l_g^* - \theta_b l_b^*) + (1 - \beta)k^*.
\]

The government can select any imaginary allocation that satisfies the fooling and incentive compatibility constraints, say \((c^I_m, k^I_m)_{m \in \Theta}\). It can proceed to pin down the imaginary savings subsidy

\[
1 + \tau^I_m = \frac{w'(k^I_m)}{u'(c^I_m)}.
\]
Using the savings subsidy, it can easily find the imaginary lump-sum taxes

$$\Lambda^I_m = \theta_m l^*_m - \left( c^I_m + (1 + \tau^I_m) k^I_m \right).$$

As a result, type $\theta_m$ agent faces the following policy menu: \[ \left\{ \left( \tau^I_m, \Lambda^I_m \right); \left( \tau^*, \Lambda^*_m \right) \right\}. \]

From the derivation above, the implementation typically involves non-linear savings subsidies, $\tau^I_m \neq \tau^*$. Such non-linearities are already prevalent in the present tax system. For example, a feature of IRA and 401(k) accounts is that annual contributions are capped: rate of return below cap is higher than the rate of return above the cap, which creates a non-linear intertemporal budget constraint. The model suggests that more elaborate or complicated retirement savings tax systems may improve both consumption smoothing and insurance. For example, if the government selects the imaginary allocations such that $c^I_m < c^*$ and $k^I_m > k^*$, then it is possible that the rate of return below a certain cap is lower than the rate of return above the cap. More importantly, the model suggests that savings subsidy that differs for each productivity level can help screen the agents.

The implementation can also utilize insights from Thaler and Benartzi (2004) by exploiting the tendency for agents to exhibit status quo bias. The same behavioral bias that results in inadequate savings can cause procrastination, which leads to inertia or status quo bias. In the context of my model, the default taxes can be set at $\left( \tau^*, \Lambda^*_m \right)$, with the option of changing to $\left( \tau^I_m, \Lambda^I_m \right)$ post-reporting for type $\theta_m$ agent. Though exploiting the possible status quo bias is not needed in my model, it makes sense to exploit the status quo bias and set the default to $\left( \tau^*, \Lambda^*_m \right)$ for implementation.

Finally, notice that this implementation adheres to the rules of libertarian paternalism because the freedom of choice is not compromised in this setup. The agents are allowed to choose to consume at the imaginary allocations, but would not. A detailed discussion of paternalism is provided in Section 8.

### 6.2 Savings with Sophisticates

For sophisticated agents, the government provides a menu of real and threat allocations. Similar to non-sophisticated agents, the problem is analogous to the general model with preference changes in Section 4. Also, since Assumption 1 and 2 are trivially satisfied, the

---

7 With linear tax instruments, the agents must equate their marginal rate of intertemporal substitution to the same price before the taste change and after the taste change. As a result, linear instruments do not allow the government to separate the screening problem from the consumption smoothing problem, so it restricts the extent of insurance the government is able to provide.

8 Sunstein and Thaler (2008) have argued for ‘libertarian paternalist’ policies that would ‘nudge’ individuals to choosing the appropriate course of action without compromising the freedom of choice.
government can achieve first best welfare.

**Corollary 4** It is optimal to use a threat mechanism. The optimal allocation for the environment with private information and sophisticated time-inconsistent agents is the same as the allocation in the environment without private information.

To demonstrate Corollary 4, consider the setting with two productivity types and a utilitarian government introduced previously. The incentive compatibility constraints are

\[
\begin{align*}
u (c^R_g) - h(l^R_g) + w(k^R_g) &\geq u(c^T_b) - h \left( \frac{\theta_b l^T_b}{\theta_g} \right) + w(k^T_g), \\
u (c^R_b) - h(l^R_b) + w(k^R_b) &\geq u(c^T_g) - h \left( \frac{\theta_g l^T_g}{\theta_b} \right) + w(k^T_g),
\end{align*}
\]  

and the credible threat constraints are

\[
\begin{align*}
u (c^R_g) - h(l^R_g) + \beta w(k^R_g) &\geq u(c^T_b) - h(l^T_g) + \beta w(k^T_g), \\
u (c^R_b) - h(l^R_b) + \beta w(k^R_b) &\geq u(c^T_g) - h(l^T_g) + \beta w(k^T_g), \\
u (c^T_b) - h \left( \frac{\theta_b l^T_b}{\theta_g} \right) + \beta w(k^T_b) &\geq u(c^R_g) - h \left( \frac{\theta_g l^T_g}{\theta_b} \right) + \beta w(k^T_g), \\
u (c^T_g) - h \left( \frac{\theta_g l^T_g}{\theta_b} \right) + \beta w(k^T_g) &\geq u(c^R_b) - h \left( \frac{\theta_b l^T_b}{\theta_g} \right) + \beta w(k^T_b),
\end{align*}
\]  

In a Mirrlees setting, the high productivity agent has an incentive to pretend to be a lower productivity agent to decrease labor supply while enjoying the gains of full insurance. Hence, in a relaxed problem, the threat is targeted at the high productivity type to discourage him from pretending to be the low productivity type. I will focus on the implications of \(18\), \(21\) and \(22\), while ignoring \(19\), \(20\) and \(23\).

By Corollary 4, \((c^R, R^R, y^R)\) will be the first best allocation. The first best allocation has consumption and saving equated across types and perfect smoothing across periods, and it has the efficient agents working more than the inefficient agents. In Figure 4, the credible threat constraint narrows down the possible threat allocation \((c^T_b, l^T_b)\).

The threat allocation, \((c^T_b, k^T_b, l^T_b)\), needs to satisfy the incentive compatibility constraint \((18)\). Let \(\Phi^i_{j,k} = u(c^j) - h \left( \frac{y/\theta_k}{\theta_j} \right)\) and \(\Phi^i_{k,k} = \Phi^i_k\), then from \((18)\) and \((22)\), and from the fact that \(\theta_g > \theta_b\) and \(\beta < 1\), the first best and threat allocations have to satisfy

\[
\Phi^T_{b,g} > \Phi^R_{b,g} > \Phi^R_g > \Phi^T_g,\text{ and } k^R > k^T_b.
\]
Figure 3: Finding the Threat Allocation: Part I

Figure 3 shows how the incentive compatibility constraint restricts the set of threat allocations. The steeper solid (red) curve represents the indifference curve of the ex-post utility for the efficient agent who pretends to be inefficient. The flatter solid (blue) curve represents the indifference curve of the ex-ante utility for the efficient agent who reports truthfully. The dashed (red) curve represents the indifference curve of the ex-post utility for the truthful inefficient agent. Figure 3 shows that the government can choose \((c^T_b, k^T_b, l^T_b)\) such that the incentive compatibility constraint (18) is satisfied by decreasing \(k^T_b\) and increasing \(\Phi^T_{b,g}\).

By Assumption 2, the government can increase \(\Phi^T_{b,g}\) without violating the credible threat constraints (21) and (22). To demonstrate this, let \(\Gamma^R = u(c^R) + \beta w(k^R)\) and \(\Gamma^T_b = u(c^T_b) + \beta w(k^T_b)\), and from (21) and (22) and by the convexity of \(h(\cdot)\) and \(\theta_g > \theta_b\), the following must hold

\[
\Gamma^T_b > \Gamma^R, \text{ and } l^T_b > l^R. \]

Figure 4 shows how the threat allocation can be chosen to satisfy the credible threat constraint. In Figure 4, the flatter solid (blue) curve represents the indifference curve of the ex-post utility for the efficient agents at allocation \((c^R, k^R, y^R_g)\). The steeper solid (red) curve represents the indifference curve of the ex-post utility for the inefficient agents at allocation \((c^R, k^R, y^R_b)\). The dashed (blue) curve represents the indifference curve of the ex-post utility for the efficient agent at allocation \((c^R, k^R, y^R_b)\), in essence, if the productive agent misreported his productivity type as inefficient and chose the real allocations. The
threat allocation \((c^T_b, k^T_b, l^T_b)\) has to be in the area bounded by the steeper solid (red) curve and the dashed (blue) curve in the north-east region.

The key to making a threat credible and incentive compatible is to decrease savings and increase present utility from choosing the threat allocation. The ex-post utility has a present bias, so a threat allocation would seem more appealing than the real allocation if it indulges the bias. To make sure that only a misreporting agent is punished, the government exploits the fact that the efficient agent has lower marginal rate of substitution in consumption and output than less efficient agents. Hence, the threat allocation can always be chosen such that only a misreporting agent would fall for it.

### 6.2.1 Implementation: Loans and Commitment Plans

With sophisticated agents, the government needs to be explicit with the provision of a savings plan. At the same time, the policy needs to screen the productivity types of the agents. Consider the two productivity types case with an utilitarian government. The government wishes to implement the first best allocation \((c^*_m, k^*_m, l^*_m)_{m \in \Theta}\). Recall that the first best allocation has perfect consumption smoothing, full insurance across types and efficient provision of labor for all productivity types, and it satisfies the government budget constraint. To implement the first best allocation, the government needs to choose a set of threat allocations for the inefficient type to deter inefficient provision of labor from the productive agents.
Similar to the fooling mechanism, the real savings subsidy is chosen to smooth consumption across periods and is the same for all skill levels

\[ 1 + \tau_g^* = 1 + \tau_b^R = \beta. \]

Also, the transfers have to satisfy the government budget constraint and are the same as with the fooling: for the high productivity agents,

\[ \Lambda_g^* = \pi_b (\theta_g l_g^* - \theta_b l_b^*) + (1 - \beta) k^*, \]

and for the low productivity agents,

\[ \Lambda_b^R = -\pi_g (\theta_g l_g^* - \theta_b l_b^*) + (1 - \beta) k^*. \]

Next, the government needs to choose the threat policy, \((\tau_b^T, \Lambda_b^T)\), to deter the efficient agents from pretending to be inefficient.

The threat policy needs to undo the consumption smoothing. The government can do this by choosing \(\tau_b^T > 0\), in essence, a savings tax. This is pinned down by choosing a low \(k_b^T\) and a high \(c_b^T\), so

\[ 1 + \tau_b^T = \frac{w'(k_b^T)}{u'(c_b^T)}. \]

Simultaneously, it needs to make sure that the inefficient agents do not choose \(\tau_b^T\). Therefore, if the agent chooses \(\tau_b^T\), then the government can require the agent to forfeit the positive transfers, \(\Lambda_b^T = 0\), and face an income tax instead. To choose the appropriate marginal income tax for the threat policy \((MTR_b^T)\), the government can first choose \(l_b^T\) and \(c_b^T\) such that the real inefficient agent would never choose it but a misreporting efficient agent would. As a result, the government has the marginal income tax as

\[ MTR_b^T = 1 - \frac{\theta_b h' \left( \theta_b l_b^T / \theta_g \right)}{\theta_g u'(c_b^T)}. \]

As a result, efficient agent faces the following menu of linear savings subsidy and lump-sum transfer: \((\tau_g^*, \Lambda_g^*)\). The inefficient agent faces the following menu: \(\{(\tau_b^R, \Lambda_b^R), (\tau_b^T, \Lambda_b^T, MTR_b^T)\}\).

The derivation above suggests that the optimal savings subsidy for the efficient agent can be linear, as opposed to the non-linear policy designed to fool the non-sophisticated agents. The difficulty is in choosing the appropriate instruments to threaten misreporting agents.
7 Model with Time Consistent Agents

In the previous sections, I assumed that all agents were time-inconsistent. This is an extreme assumption. If time-inconsistent agents are fully naïve, then the government cannot screen for time-inconsistency, since all agents share the same belief.

The extent of the fooling is limited by the presence of time consistent agents because the allocation used to elicit truth-telling would also be consistent with the realized consumption and savings. This places a natural limit on the imaginary allocations, whereas in the benchmark model, the imaginary allocations do not need to satisfy the government budget constraint.

The government is uncertain whether the agents are time consistent ($\beta = 1$) or time-inconsistent ($\beta < 1$). If the agents are time-inconsistent, then they are also fully naïve, so they are susceptible to being fooled by the government. The prior belief of the government that the agents are time-inconsistent is $\Pr(\beta < 1) = \phi < 1$. For expositional purposes, the probability of being time-inconsistent is independent from the agent’s production capabilities.

I will focus on the model with consumption and savings. The rest of the setup is the same as the previous section with two productivity types, $\Theta = \{\theta_b, \theta_g\}$ and a utilitarian social welfare function. The government policy in terms of allocation is also the same as in the previous sections, with both imaginary and real allocations. All agents would evaluate the merits of truth-telling using the imaginary allocations, but the difference now being a time consistent agent would not choose the real allocations when they make their intertemporal decision.

Recall that in Section 3 the imaginary allocations did not have to satisfy the government budget constraint, because the government was certain that a preference change would occur. However, if the agents are time consistent, the government will fail to successfully deceive them and, if the agents chose the imaginary allocations, the government budget constraint will be violated. The government budget constraint is

$$\phi \left[ \sum_{\theta_m \in \Theta} \pi_m (\theta_m l_m - c_m^R - k_m^R) \right] + (1 - \phi) \left[ \sum_{\theta_m \in \Theta} \pi_m (\theta_m l_m - c_m^I - k_m^I) \right] = 0. \quad (24)$$

The government maximizes the utilitarian social welfare function subject to the incentive compatibility constraints (12) and (13), the fooling constraints (14), (15), (16) and (17) and the government budget constraint (24).

As was mentioned, the presence of time consistent individuals limits the government’s ability to deceive. Moreover, it is not optimal to fool the least productive agent.

**Lemma 2** The government fools the efficient agent, but does not fool the inefficient agent.
To see the intuition for Lemma 2, first notice that information asymmetry would affect the allocations. Since the time-inconsistent agents are fully naïve, both the time consistent and time-inconsistent agents would use the same imaginary allocations to evaluate their reporting strategy. However, the lack of preference switching for time consistent individuals means that their planned consumption is consistent with their actual consumption. As a result, the incentive compatibility constraint for the productive type would bind. If not, then the government can improve the welfare of the time consistent bad type agents.

Even though insurance is limited by the presence of time consistent individuals, the government is able to decrease the gap in welfare between productivity types by fooling the time-inconsistent productive agents. However, Lemma 2 says that fooling the least productive agents does not help bridge the gap. The intuition for this result can be seen from the following indifference curves for consumption and savings in Figure 5.

![Figure 5: Limited Fooling with Time Consistent Agents](image)

In Figure 5, by fooling the productive agents, the government can bridge the gap in utilities between the two types of agents. However, if the government fools the least productive agent, the government is unable to achieve full insurance without violating the incentive compatibility constraint.

Lemma 2 implies that the agents can be divided into three groups and their allocations are \( \{(c_b, k_b, l_b); (c_g^I, k_g^I, l_g); (c_g^R, k_g^R, l_g)\} \), where both time-inconsistent and time consistent low productivity agents are pooled together. The following proposition characterizes the optimal allocation.

31
Proposition 4 The optimal allocation \( \{(c_b, k_b, l_b); (c_g^I, k_g^I, l_g^I); (c_g^R, k_g^R, l_g^R)\} \) has the following properties

i. The inefficient agents smooth consumption over time optimally. In essence, \( u'(c_b) = w'(k_b) \).

ii. The time consistent efficient agents save too much and the naïve time-inconsistent inefficient agents have inadequate savings. In essence, \( u'(c_g^I) > w'(k_g^I) \) and \( u'(c_g^R) < w'(k_g^R) \).

iii. The inefficient agents and the time consistent efficient agents produce too little. In essence, \( u'(c_b) > \frac{1}{\theta_b} h'(l_b) \) and \( u'(c_g^I) > \frac{1}{\theta_g} h'(l_g) \).

iv. The naïve time-inconsistent efficient agents produce too much. In essence, \( u'(c_g^R) < \frac{1}{\theta_g} h'(l_g) \).

v. The savings of each type of agent has the following relationship: \( k_g^I > k_g^R > k_b \).

vi. The consumption of each type of agent has the following relationship: \( c_g^R > c_g^I \) and \( c_b > c_g^R \).

In terms of social welfare, Corollary 3 shows how the government can achieve full insurance without sacrificing efficiency when \( \phi = 1 \). The usual Mirrlees taxation case with time consistent agents occurs when \( \phi = 0 \), and the second best welfare occurs. It is natural to presume that social welfare increases as the measure of naïve time-inconsistent individuals increases. The following proposition confirms this.

Proposition 5 The utilitarian social welfare increases as the proportion of fully naïve time-inconsistent individuals increases.

8 Discussions

In this section, I will address some concerns regarding the assumptions in the paper and some of the implications of the results, in particular, the message of paternalism this paper seems to imply.

8.1 Impediments to Implementation

This paper has focused on a setting where the government can screen time-inconsistent agents without impunity. In the last section, I showed how the presence of time consistent
agents would impede the screening process and make distortions necessary. Here, I will discuss other situations where the first best may not be achievable.

### 8.1.1 Political Economy

In a political economy, the incentives to be re-elected would constrain the set of implementable policies. Even for benevolent political candidates, if the primary goal is to win the election, political incentives would distort the choice of policies.

The intertemporal rate of consumption could be distorted. This is especially true when elections are held after the onset of the present bias. Suppose an election occurs after the agents’ preferences change, then the political candidates have an incentive to undo policies that encourage savings. The competition for votes could force the candidates to pander to the voters’ desire for present consumption and undermine the implementation of optimal savings policies. The timing of elections has shown to be of crucial importance in models with time-inconsistent voters. Bisin, Lizzeri and Yariv (2014) showed how political candidates would exploit the voters’ present bias and undo the incentives for private commitment when elections are held in tandem with the intertemporal decisions of the agents. This fact is true regardless of the agent’s cognitive type.

### 8.1.2 Outside Commitment Devices

People with self-control problems can choose from a wide array of commitment devices available in the market, for example, illiquid assets. There is also a rising market demand for commitment devices. For example, StickK, Pact and Beeminder are some recent websites that offer contracts contingent on the completion of stated goals.

In the case of sophisticated agents, if commitment devices are available and its usage is unobservable, then threats become much less potent. This is because an agent can purchase illiquid assets and bind himself to an intertemporal allocation. This reduces the effectiveness of a threat, because the ex-post utility is maximized over a smaller consumption set. Therefore, private information could matter when commitment devices for sophisticated or partially naïve agents are available.

However, fooling mechanisms for non-sophisticated agents could override the demand for commitment devices. The government can always choose imaginary allocations that make buying an outside commitment device undesirable.
8.2 Naïveté and Non-common Priors

In the paper, I considered an economy populated by partially and diversely naïve agents. This is in sharp contrast to the existing literature on time-inconsistent preferences, which usually adopts the view that the agent is sophisticated. A partially naïve agent is not fully aware of his time-inconsistency, while a sophisticated agent is. The paper departs from the usual assumption in cognitive ability due to recent developments in psychology and behavioral economics.

DellaVigna and Malmendier (2006) studied gym membership data. They found those who chose to be members attended the gym so seldom and irregularly that they would have been better off going as non-members. This empirical phenomenon is difficult to explain with rational or even sophisticated agents. The literature has interpreted this result as evidence in support of naïveté. The gym members hold a false belief that their willingness to exercise in the present will persist in the future, which leads them to make an incorrect contracting choice. Many other papers have demonstrated such naïveté using empirical data, including an examination of the credit card market by Ausubel (1999) and Shui and Ausubel (2005). Models of partial naïveté also help explain the impact of the status quo in 401(k) plan choices, which is called the default effect. Madrian and Shea (2001) have documented the default effect on contribution rates in 401(k)s. More recently, there is also experimental data in support of naïveté. For example, Hey and Lotito (2009) have found that subjects display dynamically inconsistent behavior in-line with naïveté.

A common objection to the adoption of the partial naïveté assumption is that agents have the ability to learn. After repeated decision making, an agent should and is expected to learn about his behavioral bias and thus, become fully aware of his time-inconsistency. He may even correct it accordingly. However, on the issue of retirement, most people retire only once in their lifetime. Therefore, it is safe to assume that people are unable to learn about their time-inconsistency when it comes to retirement decisions, and remain largely unaware of their behavioral bias. Also, on the issue of retirement decisions, recent scientific discoveries have shown that ageing can have detrimental effects on parts of the brain responsible for financial planning. This includes poorer memory and ability to imagine possible future scenarios.

There is also evidence that people do a poor job of learning about their future preferences and thus remain ignorant of their time-inconsistency problem even after repeated decision making. The psychology literature has identified several possible forces that obstruct learn-

---

9 For example, findings have shown reductions in the hippocampal volume and deterioration of the prefrontal cortex due to ageing, which impedes the ability to recall past experiences and make reasoned judgments. For a detailed account on how ageing could affect financial decision making, see Weierich et al. (2011).
ing. For example, it is recognized that we tend to disregard information that runs counter to our beliefs, while paying much closer attention to information that could support our beliefs. This is called confirmation bias. Another related phenomenon documented is conservatism, which describes an updating bias where individuals give too much credence to past observations and not enough weight to new information. Another psychological phenomenon that could obstruct learning is the fact that human memory often displays limitations, so information updating is not performed on the full set of signals. 10

I also implicitly assumed that while the agents were partially naïve, the government can anticipate the change in discount factors correctly, which creates the conflict in beliefs. I believe this to be a reasonable assumption. The government has access to all agents’ saving behavior in the economy, while the agent has limited knowledge of this. Also, the government employs researchers, such as experts at the Bureau of Labor Statistics, studying the savings behavior of its agents. Therefore, it is safe to assume that the government is better informed about the agents’ systematically changing time preferences.

### 8.3 Alternative Welfare Criteria

The choice of the welfare criteria in a multi-selves model is often left to the modeler’s own discretion. In line with most of the work in this literature, I chose maximizing the ex-ante utility of the agents as the government’s welfare objective. This view is motivated by the fact that agents wish and plan to consume allocations according to their ex-ante utility, but are subject to the whims of their ex-post utility, which they see as falling into the hands of uncontrolled temptations. This is modeled by the fact that the agents use their ex-ante utility to evaluate the incentive compatibility constraints.

However, this does not preclude the government from placing strictly positive welfare weights on the ex-post utility. The motivation for it may be that the government hopes the agent could be more spontaneous and enjoy life while he or she is young. If the government chooses to do so, it is still able to achieve the first best allocation under the new welfare criteria. In other words, the results of the paper do not change much if the welfare criterion is different. As a result, the main idea presented is robust to subjective judgment for the appropriate welfare criterion.

Though the results of the paper do not change with regards to the welfare choice, I prefer using the ex-ante utility as the main welfare criterion. As was mentioned before, the ex-post utility reflects unreasoned and instinctive preferences that the agent inherently

---

10Gottlieb (2011) studies a model of learning with confirmation bias and conservatism and finds that learning is never complete even in the limit. Wilson (2014) studies a model with limited memory which generates imperfect learning.
wishes to avoid. This interpretation is consistent with the Bernheim and Rangel (2004) interpretation of ex-post selves. Consequently, it is natural to evaluate welfare according to the ex-ante preferences.

8.4 Paternalism

Recently, in light of developments in behavioral economics, an argument has been made for paternalistic policies that aim to aid individuals in overcoming their undesirable tendencies. Most have argued for paternalistic policies that limit the breach of sovereignty by examining mechanisms that would alter the choices of an individual with behavioral biases, while having little effect on individuals without such biases. However, this examination has been done in isolation of other goals that the government may have. In other words, a systematic analysis of how paternalism interacts with other motives, such as insurance provision, has not been discussed.

In this paper, the government has both insurance and paternalistic goals. I abstract from issues of sovereignty and examine the optimal policy without ulterior constraints. I show that the first best outcome is achievable provided that exploiting the time-inconsistency of individuals is acceptable. Though my analysis discusses both policy and welfare implications, it is not meant to be a normative analysis. The arguments for and against paternalism are equally compelling, but this paper is not meant to take a stand on either side. In fact, it could be used to argue for paternalism and for anti-paternalism. Those in favor of paternalism could interpret the results as a further validation of manipulating individuals, not only for their own good, but for increasing the social welfare. On the other hand, anti-paternalists could argue this paper shows that even a rational and benevolent government with paternalistic goals would be motivated to go too far in exploiting agents, and that the gains in social welfare come at an exorbitant price. For example, a soft paternalistic savings plan may succumb to the government’s insurance motives and trigger a slippery slope towards more intrusive paternalism, as described in Rizzo and Whitman (2009). A strong case can also be made for the moral basis of deceiving individuals who are not aware of their biases. In fact, for deception to be sustainable, the paper recommends not educating individuals about their biases.

Though this paper does not add to the discussion of whether paternalism is desirable, I believe that it does show the importance and need for rigorous analysis of paternalistic

---

11 In addition to libertarian paternalism as prescribed by Sunstein and Thaler (2008), Camerer et al. (2003) have also defended the implementation of paternalistic policies provided that they bring large benefits to boundedly rational individuals while limiting their cost on rational individuals. They call this ‘asymmetric paternalism’ since it leaves rational agents unaffected.
policies. An uninhibited analysis of paternalism helps us understand the form of the optimal policies, which leads naturally to a discourse of whether these policies should be adopted or rejected based on moral or philosophical considerations.

9 Summary and Conclusion

In this paper, I examined the optimal policies for a government facing a population of time-inconsistent agents with hidden productivity. I showed that the first best welfare is attainable despite the presence of asymmetric information. This is because with non-sophisticated agents, the government can separate types by exploiting their inability to precisely forecast the eminent taste change in the future. With sophisticated agents, the government can provide them with a commitment device and threaten to undo the commitment if they do not tell the truth about their productivity. The optimal policy requires nonlinear type-specific savings subsidies. Type specific subsidies can help the government separate the types, and the non-linearity helps deceive the non-sophisticated agents and threaten the sophisticated agents. I also discussed several settings where such a strong result would not hold.

The result presented in this paper could be applied to models of industrial organization. For example, it could be applied to a model of gym membership where consumers have heterogeneous marginal value of attending gym, but are not fully aware of their time-inconsistency. The gym can fully price discriminate with a membership contract that is type specific and includes heavily discounted usage prices in the future with an expensive alternative option. Consumers would be attracted by the discounts they would enjoy in the future, mis-predicting the fact that their tastes would change and would prefer the alternative option. The concept of fooling and issuing threats could potentially be used in a wider array of mechanism design problems with dynamically inconsistent agents.

A serious issue that is not being addressed by the present model is the lack of learning by the non-sophisticated agents. A dynamic model with non-dogmatic agents can potentially shed light on this issue. If non-sophisticated agents are expected to learn about their present bias problem, the government might need to adjust their optimal policies each period to continue to deceive the agents. Once the agents fully realize the extent of their bias, the government switches to the threat mechanism. However, if the rate of learning is hidden from the government, then the appropriate degree of fooling or threat is unknown to the government.

Another serious issue that was not rigorously addressed in this paper was the desirability of paternalist regulations. This paper discusses paternalist policies without regards to other
important issues, such as the morality and sovereignty concerns involved with the government attempting to fool and issuing threats to individuals. This paper demonstrates the need for systematic and constructive analysis of such issues, since the adoption of certain welfare measures combined with paternalist measures could potentially lead to the enactment of undesirable policies.

Appendix A: Proofs

Proof of Proposition 1

Let $\mu_I^m$ ($\mu_R^m$) be the Lagrange multiplier on the fooling constraint for productivity type $\theta_m$ to preferring the imaginary (real) allocation over the real (imaginary) allocation. Finally, let $\lambda(\theta_m'; \theta_m)$ be the Lagrange multiplier for the incentive compatibility constraint on type $\theta_m$ misreporting to be $\theta_m'$.

Let us begin with magnitude naiveté, and analyze the first order conditions for the imaginary consumption, $\forall \theta_m \in \Theta$ and $\forall n \in N$,

$$\left\{ \sum_{\theta_m' \in \Theta} \left[ \lambda(\theta_m'; \theta_m) - \lambda(\theta_m; \theta_m') \right] + \alpha \mu_I^m \right\} \frac{\partial U}{\partial c_{m,n}^I} = \left[ \mu_R^m - (1 - \alpha) \mu_I^m \right] \frac{\partial V}{\partial c_{m,n}^I}. $$

The marginal utility of consumption can be taken out of the summation due to the assumption of separability in labor and consumption. By Assumption I and the fact that $\lim_{c_n \to 0} \frac{\partial V}{\partial c_n} = +\infty$ and $\lim_{c_n \to 0} \frac{\partial V}{\partial c_n} = +\infty$, so consumption is strictly positive (non-negativity constraints never bind), the following is immediate

$$\sum_{\theta_m' \in \Theta} \left[ \lambda(\theta_m'; \theta_m) - \lambda(\theta_m; \theta_m') \right] + \alpha \mu_I^m = \mu_R^m - (1 - \alpha) \mu_I^m = 0. $$

Summing across all types yields

$$\sum_{\theta_m \in \Theta} \sum_{\theta_m' \in \Theta} \left[ \lambda(\theta_m'; \theta_m) - \lambda(\theta_m; \theta_m') \right] = 0 \quad \text{(25)}$$

This implies that $\alpha \sum_{\theta_m \in \Theta} \mu_I^m = 0$. As a result, the Kuhn-Tucker necessary conditions imply $\mu_I^m = 0$, which gives us $\mu_R^m = 0$.

For frequency naiveté, the first order conditions for the imaginary consumption, $\forall \theta_m \in \Theta$
and \( \forall n \in N, \)

\[
\left\{ \alpha \sum_{\theta_{m'} \in \Theta} \left[ \lambda(\theta_{m'}; \theta_m) - \lambda(\theta_m; \theta_{m'}) \right] + \mu^L_m \right\} \frac{\partial U}{\partial c^L_{m,n}} = \mu^R_m \frac{\partial V}{\partial c^L_{m,n}}.
\]

By Assumption 1, the following must hold \( \alpha \sum_{\theta_{m'} \in \Theta} \left[ \lambda(\theta_{m'}; \theta_m) - \lambda(\theta_m; \theta_{m'}) \right] + \mu^L_m = \mu^R_m = 0. \)

Using a similar method as in (25), it follows that \( \mu^L_m = \mu^R_m = 0 \) for all productivity types.

Since \( \mu^L_m = \mu^R_m = 0, \) the first order conditions on the imaginary consumption for both types of naïveté and for all \( \theta_m \in \Theta \) have the following property

\[
\sum_{\theta_{m'} \in \Theta} \lambda(\theta_{m'}; \theta_m) = \sum_{\theta_{m'} \in \Theta} \lambda(\theta_m; \theta_{m'})
\]  

Consider the most productive agent \( \theta_M \) and assume that there exists a type \( \theta_{\hat{m}} \) such that \( \lambda(\theta_{\hat{m}}, \theta_M) > 0. \) By (26), there exists a type \( \theta_{\hat{m}} \) such that \( \lambda(\theta_M, \theta_{\hat{m}}) > 0. \) In other words, if the most productive type is indifferent between truth-telling and pretending to be a less efficient type \( \theta_{\hat{m}}, \) then there is another type of agent \( \theta_{\hat{m}} \) that would be indifferent between truth-telling and pretending to be the most efficient type.

By Assumption 2, there exists another allocation for type \( \theta_M \) with larger \( c^I(\theta_M) \) and more labor \( l(\theta_M) \) such that type \( \theta_M \) strictly prefers it to the original one and type \( \theta_{\hat{m}} \) would never choose this new allocation. To see this, let \( (y(\theta_M), c^I(\theta_M)) \) and \( (y(\theta_{\hat{m}}), c^I(\theta_{\hat{m}})) \) denote the original allocations, and \( (y^*(\theta_M), c^I(\theta_M)) \) and \( (y^*(\theta_{\hat{m}}), c^I(\theta_{\hat{m}})) \) the new allocations. Choose a good \( n \in N \) such that its production depends on labor. Let \( MRS(y, c)_{\lambda,n} \) and \( MRS(y, c)_{\lambda,n} \) be the marginal rate of substitution of the two types, \( \theta_M \) and \( \theta_{\hat{m}}, \) in \( y_n \) and \( c_n. \) The easiest way to construct the new allocations is to choose it such that \( (y^*_n(\theta_M), c^I(\theta_{\hat{m}}) + \epsilon, c^I(\theta_M) + \omega e) \) and \( (y^*_n(\theta_{\hat{m}}), c^I(\theta_{\hat{m}}) + \omega e) \) are the new allocations, where \( MRS(y(\theta_M), c^I(\theta_M)) < \omega < MRS(y(\theta_{\hat{m}}), c^I(\theta_{\hat{m}})) \) and \( \epsilon \) is chosen to be sufficiently large so that type \( \theta_{\hat{m}} \) is strictly worse off when he pretends to be type \( \theta_M. \) Since the imaginary allocation does not enter the government’s welfare criterion, the extra output of \( \epsilon \) can then be redistributed, which raises the social welfare. Therefore, \( \lambda(\theta_M, \theta_{\hat{m}}) = 0 \) for all \( \theta_{\hat{m}} \) which contradicts (26), so \( \lambda(\theta_{\hat{m}}, \theta_M) = 0 \) for all \( \theta_{\hat{m}}. \)

The same argument can be repeated for all lower productivity types. In essence, it is never optimal for \( \lambda(\theta_m, \theta_{m'}) > 0 \) when \( \theta_{m'} < \theta_m, \) so by (26), it is also not optimal for \( \lambda(\theta_m, \theta_{m'}) > 0 \) when \( \theta_{m'} > \theta_m. \) Therefore, all Lagrange multipliers for all incentive compatibility constraints are non-positive, which shows that the incentive compatibility constraints are inactive. This proves Proposition I. ■
Proof of Corollary 1

By Proposition 1, the government can implement the first best allocation. Let 
\( \{(c^R(\theta_m), l(\theta_m))\}_{\theta_m \in \Theta} \) be the first best allocation. Suppose the government does not fool the agents, then by definition, for all \( \theta_m \in \Theta \), 
\( (c^R(\theta_m), l(\theta_m)) = (c^I(\theta_m), l(\theta_m)) \), which violates the incentive compatibility constraint for some types. It follows that for some types, 
\( (c^R(\theta_m), l(\theta_m)) \neq (c^I(\theta_m), l(\theta_m)) \), and thus the government must implement a fooling mechanism to achieve the first best allocation.

Proof of Proposition 2

Let \( \mu^T_m(\theta_m'; \theta_m) \) be the Lagrange multiplier on the credible threat constraint for productivity type \( \theta_m \) to preferring the threat allocation over the real allocation when reporting as type \( \theta_m' \). Let \( \mu^R_m \) be the Lagrange multiplier on the credible threat constraint for productivity type \( \theta_m \) to preferring the real allocation over the threat allocation when reporting truthfully. Let \( \lambda(\theta_m'; \theta_m) \) be the Lagrange multiplier for the incentive compatibility constraint on type \( \theta_m \) misreporting to be \( \theta_m' \).

The first order necessary conditions for the threat consumption, \( \forall \theta_m \in \Theta \) and \( \forall n \in N \), is
\[
\left[ \sum_{\theta_m' \in \Theta} \lambda(\theta_m; \theta_m') \right] \frac{\partial U}{\partial c^T_{m,n}} = \left[ \sum_{\theta_m' \in \Theta} \mu^T(\theta_m; \theta_m') - \mu^R_m \right] \frac{\partial V}{\partial c^T_{m,n}}
\]

The marginal utility of consumption can be taken out of the summation due to the assumption of separability in labor and consumption. By Assumption 1 and the fact that \( \lim_{c_n \to 0} \frac{\partial U}{\partial c_n} = +\infty \) and \( \lim_{c_n \to 0} \frac{\partial V}{\partial c_n} = +\infty \), so consumption is strictly positive (non-negativity constraints never bind), the following is result is apparent
\[
\sum_{\theta_m' \in \Theta} \mu^T(\theta_m; \theta_m') - \mu^R_m = \sum_{\theta_m' \in \Theta} \lambda(\theta_m; \theta_m') = 0.
\]

As a result, the Kuhn-Tucker necessary conditions imply \( \lambda(\theta_m; \theta_m') = 0 \), for all \( \theta_m, \theta_m' \in \Theta \).

To check that the credible threat constraints do not distort the allocations, suppose \( \mu^R_m > 0 \), then there exists a \( \theta_m' \) such that \( \mu^T(\theta_m; \theta_m') > 0 \). Therefore, by the complementary slackness conditions, the following must hold
\[
V(c^R(\theta_m), l^R(\theta_m)) = V(c^T(\theta_m), l^T(\theta_m)),
\]
\[
V \left[ c^T(\theta_m), G(y^T(\theta_m), x^T(\theta_m); \theta_m') \right] = V \left[ c^R(\theta_m), G(y^R(\theta_m), x^R(\theta_m); \theta_m') \right].
\]

For \( \theta_m' > \theta_m \), by Assumption 2 both the output and consumption of the threat allocation
can be increased such that only the relatively more efficient agent will choose it. (Adopting a similar process in the proof of Proposition 1.) Hence, the threat allocation can be chosen such that \( \mu^R_m = \mu^T(\theta_m; \theta_{m'}) = 0, \forall \theta_m, \theta_{m'} \in \Theta \). Similarly, for \( \theta_m > \theta_{m'} \), by Assumption 2, the threat allocation can be chosen such that \( \mu^R_m = \mu^T(\theta_m; \theta_{m'}) = 0, \forall \theta_m, \theta_{m'} \in \Theta \) by decreasing the output and consumption of the threat allocation. This completes the proof.

Proof of Corollary 2

First notice that by Assumption 2, the convex combination \( W = \alpha U + (1 - \alpha)V \) also satisfies the single crossing condition. For magnitude naïveté of sophistication level \( \alpha \), the first order necessary conditions for the threat consumption, \( \forall \theta_m \in \Theta \) and \( \forall n \in N \), is

\[
- \left\{ \alpha \left[ \sum_{\theta_{m'} \in \Theta} \mu^T(\theta_m; \theta_{m'}) - \mu^R_m \right] - \sum_{\theta_{m'} \in \Theta} \lambda(\theta_m; \theta_{m'}) \right\} \frac{\partial U}{\partial c^T_{m,n}} = \eta_m
\]

\[
= \left\{ (1 - \alpha) \left[ \sum_{\theta_{m'} \in \Theta} \mu^T(\theta_m; \theta_{m'}) - \mu^R_m \right] - \mu^R_m \right\} \frac{\partial V}{\partial c^T_{m,n}},
\]

where \( \eta_m \) is the Lagrange multiplier on (11). Following a similar argument in the proof of Proposition 2, it must be that

\[
\sum_{\theta_{m'} \in \Theta} \mu^T(\theta_m; \theta_{m'}) - \mu^R_m = \sum_{\theta_{m'} \in \Theta} \lambda(\theta_m; \theta_{m'}) = \eta_m = 0.
\]

The result for magnitude naïveté then immediately follows by following the argument outlined for the proof of Proposition 2. For frequency naïveté, replace \( V \) with \( (1 - \alpha)V \) in the credible threat constraints (9) and (10), and by replacing the incentive compatibility constraint with

\[
U \left( c^R(\theta_m), l^R(\theta_m) \right) \geq \alpha U \left[ c^R(\theta_{m'}), G \left( y^R(\theta_{m'}), x^R(\theta_{m'}); \theta_m \right) \right] + (1 - \alpha) U \left[ c^T(\theta_m), G \left( y^T(\theta_{m'}), x^T(\theta_{m'}); \theta_m \right) \right],
\]

for all \( \theta_m, \theta_{m'} \in \Theta \). The first order conditions for the threat allocations remain the same as the one in the proof for Proposition 2, and the result follows.

Proof of Lemma 1

First, look at fooling mechanisms for magnitude naïveté. If a fooling mechanism

\[
\left\{ (c^R(\theta_m), c^I(\theta_m)) ; l(\theta_m), x(\theta_m) \right\}_{\theta_m \in \Theta}
\]
is effective for an agent with sophistication level $\alpha$ of magnitude naïveté, then the following fooling constraint must be satisfied

$$\alpha U(c^I(\theta_m), l(\theta_m)) + (1 - \alpha) V(c^I(\theta_m), l(\theta_m))$$

$$\geq \alpha U(c^R(\theta_m), l(\theta_m)) + (1 - \alpha) V(c^R(\theta_m), l(\theta_m)) .$$

Since the other fooling constraint is also satisfied, for any $\alpha > \alpha$, the fooling constraint is relaxed and all other constraints are unchanged. Hence, a fooling mechanism that is effective for $\alpha$ is effective for more naïve agents.

Next, fooling mechanisms for frequency naïveté also have the same form as the mechanism above. Suppose the fooling mechanism is effective for an agent with sophistication level $\alpha$ of frequency naïveté, then the following incentive compatibility constraint must be satisfied

$$\alpha U(c^I(\theta_m), l(\theta_m)) + (1 - \alpha) U(c^R(\theta_m), l(\theta_m))$$

$$\geq \alpha U[c^I(\theta_{m'}, G(y(\theta_{m'}), \mathbf{x}(\theta_{m'}); \theta_m)] + (1 - \alpha) U[c^R(\theta_{m'}, G(y(\theta_{m'}), \mathbf{x}(\theta_{m'}); \theta_m)].$$

Due to separability between consumption and effort, the utility function can be written as

$$U(c, l) = u(c) - h(l),$$

so the incentive compatibility constraint becomes

$$\alpha [u(c^I(\theta_m)) - u(c^I(\theta_{m'}))] + (1 - \alpha) [u(c^R(\theta_m)) - u(c^R(\theta_{m'}))]$$

$$\geq h(l(\theta_m)) - h[G(y(\theta_{m'}), \mathbf{x}(\theta_{m'}); \theta_m)].$$

Since $\psi \circ U$ is strictly concave, the first best allocation would provide full insurance: $c^R(\theta_m) = c^R(\theta_{m'})$ for all $\theta_m \neq \theta_{m'}$. As a result, for any $\theta_m, \theta_{m'} \in \Theta$ such that $\theta_m > \theta_{m'}$ it must be that $c^I(\theta_m) > c^I(\theta_{m'})$ for the incentive constraint to hold. The incentive compatibility constraint holds trivially for $\theta_m < \theta_{m'}$ of any sophistication level. It is easy to see that for any agent with $\alpha > \alpha$, the incentive compatibility constraint holds as well. Since all other constraints are invariant to the sophistication level, a fooling mechanism that is effective for sophistication level $\alpha$ is effective for more naïve agents.

Finally, for threat mechanisms, the analysis will begin by analyzing frequency naïveté. If a threat mechanism

$$\{ (c^R(\theta_m), c^T(\theta_m)), (l^R(\theta_m), l^T(\theta_m)), (x^R(\theta_m), x^T(\theta_m)) \}_{\theta_m \in \Theta} .$$
is effective for an agent with sophistication level $\alpha$ of frequency naïveté, then the incentive compatibility constraint must be satisfied

$$U\left(c^R(\theta_m), t^R(\theta_m)\right) \geq \bar{\alpha} U\left[c^R(\theta_{m'}), G\left(y^R(\theta_{m'}), x^R(\theta_{m'}); \theta_m\right)\right]$$

$$+ (1 - \bar{\alpha}) U\left[c^T(\theta_{m'}), G\left(y^T(\theta_{m'}), x^T(\theta_{m'}); \theta_m\right)\right].$$

Note that for $\theta_m > \theta_{m'}$ to implement the first best allocation, it must be that

$$U\left[c^R(\theta_{m'}), G\left(y^R(\theta_{m'}), x^R(\theta_{m'}); \theta_m\right)\right] > U\left[c^T(\theta_{m'}), G\left(y^T(\theta_{m'}), x^T(\theta_{m'}); \theta_m\right)\right].$$

If not, then the incentive compatibility constraint would not hold. As a result, for any agent with $\alpha < \bar{\alpha}$, the incentive compatibility constraint would hold as well. The incentive compatibility constraint would hold trivially for $\theta_m < \theta_{m'}$ of any sophistication level. The other constraints do not depend on $\alpha$, so a threat mechanism that is effective for sophistication level $\bar{\alpha}$ would also be effective for more sophisticated agents.

A threat mechanism for magnitude naïveté agents with sophistication level $\bar{\alpha}$ would also take the same form as the threat mechanism for frequency naïveté. The credible threat constraints are

$$\bar{\alpha} U\left[c^R(\theta_m), t^R(\theta_m)\right] + (1 - \bar{\alpha}) V\left(c^R(\theta_m), t^R(\theta_m)\right)$$

$$\geq \bar{\alpha} U\left(c^T(\theta_m), t^T(\theta_m)\right) + (1 - \bar{\alpha}) V\left(c^T(\theta_m), t^T(\theta_m)\right),$$

$$\bar{\alpha} U\left[c^T(\theta_{m'}), G\left(y^T(\theta_{m'}), x^T(\theta_{m'}); \theta_m\right)\right] + (1 - \bar{\alpha}) V\left[c^T(\theta_{m'}), G\left(y^T(\theta_{m'}), x^T(\theta_{m'}); \theta_m\right)\right]$$

$$\geq \bar{\alpha} U\left[c^R(\theta_{m'}), G\left(y^R(\theta_{m'}), x^R(\theta_{m'}); \theta_m\right)\right] + (1 - \bar{\alpha}) V\left[c^R(\theta_{m'}), G\left(y^R(\theta_{m'}), x^R(\theta_{m'}); \theta_m\right)\right].$$

Note that constraint (11) ensures that for any $\alpha < \bar{\alpha}$, a truth-telling agent would always prefer the real allocations over the threat allocations. Separability between consumption and effort helps check that the other credible threat constraint (a misreporting agent would always choose the threat allocation over the real allocation) works for more sophisticated agents as well. Let

$$V(c, l) = v(c) - h(l),$$

so the credible threat constraint can be rewritten as

$$\bar{\alpha} \left[u(c^R(\theta_{m'})) - u(c^T(\theta_{m'}))\right] + (1 - \bar{\alpha}) \left[v(c^R(\theta_{m'})) - v(c^T(\theta_{m'}))\right]$$

$$\leq h\left[G(y^R(\theta_{m'}), x^R(\theta_{m'}); \theta_m\right] - h\left[G(y^T(\theta_{m'}), x^T(\theta_{m'}); \theta_m\right].$$

43
Notice that the incentive compatibility implies
\[ u(c^R(\theta_m)) - u(c^T(\theta_m')) \geq h(l^R(\theta_m)) - h \left[ G(y^T(\theta_{m'}), x^T(\theta_{m'}); \theta_m) \right]. \]
The mechanism is effective which implies that it has full insurance, \( c^R(\theta_m) = c^R(\theta_{m'}) \), and for any \( \theta_m > \theta_{m'} \),
\[ h(l^R(\theta_m)) > h \left[ G(y^R(\theta_{m'}), x^R(\theta_{m'}); \theta_m) \right]. \]
Therefore, it must be the case that
\[ v(c^R(\theta_{m'})) - v(c^T(\theta_{m'})) < h \left[ G(y^R(\theta_{m'}), x^R(\theta_{m'}); \theta_m) \right] - h \left[ G(y^T(\theta_{m'}), x^T(\theta_{m'}); \theta_m) \right]. \]
As a result, for any agent with \( \alpha < \bar{\alpha} \), a misreporting agent would also choose the threat allocation over the real allocation. This proves that a threat mechanism that is effective for sophistication level \( \bar{\alpha} \) is also effective for more sophisticated agents. ■

**Proof of Lemma 2:**

For the proof, I will solve a relaxed problem where the incentive compatibility constraint for the least productive type is assumed to hold. Let \( \lambda_g \) denote the Lagrange multiplier on the incentive compatibility constraint for the productive type agents. The Lagrange multipliers for the other constraints follow the convention established in the previous proofs.

The first order necessary conditions for the real allocations, \( \forall \theta_m \in \Theta \), are
\[
\left( 1 + \frac{\mu_m^R - \mu_m^I}{\phi \pi_m} \right) u'(c^R_m) = \gamma,
\]
\[
\left( 1 + \frac{\mu_m^R \beta - \mu_m^I}{\phi \pi_m} \right) w'(l^R_m) = \gamma.
\]
The first order necessary conditions for the imaginary allocations are
\[
\left( 1 + \frac{\mu_g^I - \mu_g^R + \lambda_g}{(1 - \phi) \pi_g} \right) u'(c^I_g) = \gamma,
\]
\[
\left( 1 + \frac{\mu_b^I - \mu_b^R - \lambda_g}{(1 - \phi) \pi_b} \right) u'(c^I_b) = \gamma,
\]
\[
\left( 1 + \frac{\mu_g^I - \mu_g^R \beta + \lambda_g}{(1 - \phi) \pi_g} \right) w'(k^I_g) = \gamma.
\]
\[
\left(1 + \frac{\mu^I_g - \mu^R_g \beta - \lambda_g}{(1 - \phi)\pi_g}\right) w'(k^I_g) = \gamma.
\]

Furthermore, since \( \beta < 1 \), the fooling constraints imply the following relationship between real and imaginary allocations for any productivity type \( m \): \( c^R_m \geq c^I_m \) and \( k^R_m \geq k^I_m \).

I will first show that fooling the productive type is optimal. Notice that \( \lambda_g > 0 \) implies that \( \mu^R_g > 0 \), or else \( c^I_g > c^R_g \), which violates the fooling constraints. If \( \mu^I_g > 0 \), then both fooling constraints for the productive agents bind, and \( c^R_g = c^I_g \) and \( k^R_g = k^I_g \). From the first order conditions, this implies that \( \mu^R_g - \mu^I_g = \mu^R_g \beta - \mu^I_g \), which contradicts the fact that \( \mu^R_g > 0 \). Therefore, for the relaxed problem, \( \mu^I_g = 0 \) and \( \mu^R_g > 0 \), so the government fools the productive agents.

Next, to show that the low productivity agents are not fooled, notice that \( \lambda_g > 0 \) implies that \( \mu^R_b > 0 \), or else \( k^I_b < k^R_b \), which violates the fooling constraints. Given that \( \mu^I_b > 0 \), if \( \mu^R_b > 0 \), then the real allocations and the imaginary allocations are equivalent from the fooling constraints. If \( \mu^R_b = 0 \), then the first order conditions and the fooling constraints immediately imply that

\[
-\frac{\mu^I_b}{\phi \pi_b} \geq \frac{\mu^I_b - \lambda_g}{(1 - \phi)\pi_b},
\]

and

\[
-\frac{\mu^I_b}{\phi \pi_b} \leq \frac{\mu^I_b - \lambda_g}{(1 - \phi)\pi_b},
\]

so the real allocations and the imaginary allocations are equivalent. This completes the proof.

**Proof of Proposition 4:**

By Lemma 2, the government pools the low productivity agents and fools the high productivity agents. The first order necessary conditions are

\[
(\phi \pi_g + \mu^R_g)\ u'(c^R_g) = \phi \pi_g \gamma,
\]

\[
(\phi \pi_g + \mu^R_g \beta)\ w'(k^R_g) = \phi \pi_g \gamma,
\]

\[
((1 - \phi)\pi_g + \lambda_g - \mu^R_g)\ u'(c^I_g) = (1 - \phi)\pi_g \gamma,
\]

\[
((1 - \phi)\pi_g + \lambda_g - \mu^R_g \beta)\ w'(k^I_g) = (1 - \phi)\pi_g \gamma,
\]

\[
\left(1 - \frac{\lambda_g}{\pi_b}\right)\ u'(c_b) = \gamma,
\]

\[
\left(1 - \frac{\lambda_g}{\pi_b}\right)\ w'(k_b) = \gamma.
\]
\[ \left(1 + \frac{\lambda_g}{\pi_g}\right) \frac{1}{\theta_g} h'(l_g) = \gamma \]

\[ \frac{1}{\theta_b} h'(l_b) - \frac{\lambda_g}{\pi_b} \theta_g h' \left( \frac{\theta_b l_b}{\theta_g} \right) = \gamma. \]

The statement of the proposition follows.

To complete the proof, it suffices to show that in equilibrium, the incentive compatibility constraint for the least productive agent never binds.

**Proof of Proposition 5.**

Let \( W \) denote the optimal utilitarian welfare. By the envelope theorem,

\[
\frac{\partial W}{\partial \phi} = \pi_g \left[ (u(c^R_g) + w(k^R_g)) - (u(c^I_g) + w(k^I_g)) \right] + \gamma \pi_g \left[ (k^I_g - k^R_g) - (c^R_g - c^I_g) \right],
\]

where \( \gamma \) is the Lagrange multiplier on the government budget constraint. By Proposition 4, \( c^R_g > c^I_g \), and since \( u(\cdot) \) is strictly concave, it follows that

\[
u(c^R_g) - u(c^I_g) > u'(c^R_g)(c^R_g - c^I_g) = \frac{\gamma}{1 + \mu^R_g \beta \phi \pi_g} (c^R_g - c^I_g),
\]

where the last equality follows from the first order conditions on \( c^R_g \) and \( \mu^R_g \) is the Lagrange multiplier on the fooling constraints for the real allocations. Following a similar argument, the following inequality holds for savings

\[
w(k^I_g) - w(k^R_g) < \frac{\gamma}{1 + \mu^R_g \beta \phi \pi_g} (k^I_g - k^R_g).
\]

Substituting the inequalities above, the following relationship holds

\[
\frac{\partial W}{\partial \phi} > \pi_g \left[ 1 - \left( 1 + \frac{\mu^R_g}{\phi \pi_g} \right) \right] \left[ u(c^R_g) - u(c^I_g) \right] + \pi_g \left[ \left( 1 + \frac{\mu^R_g \beta}{\phi \pi_g} \right) - 1 \right] \left[ w(k^I_g) - w(k^R_g) \right]
\]

\[= \frac{\mu^R_g}{\phi} \left\{ u(c^I_g) - u(c^R_g) \right\} + \beta \left[ w(k^I_g) - w(k^R_g) \right] \}

\[= 0,
\]

where the last equality follows from the fact that the fooling constraint on the real allocations for the productive agents is binding, \( u(c^R_g) + \beta w(k^R_g) = u(c^I_g) + \beta w(k^I_g) \). This completes the proof.
Appendix B: Quasi-hyperbolic Discounting Model

I will consider a standard quasi-hyperbolic discounting model with three periods. In contrast to the model presented in the paper, the agents will work for two periods and retire at the third and final period. The analysis will focus on a direct mechanism. The within period timing will remain the same as in the paper: agents will report before the present bias occurs and then decide on labor supply along with consumption and savings decision. I will focus on the magnitude naïveté case. Since the agents make multiple savings decisions, they have an opportunity to learn about their present bias and adjust their beliefs, so this model allows the discussion of the effects of learning on the optimal policy. Following Laibson (1997) and O’Donoghue and Rabin (2001), the utility of the agents is represented as follows

\[
U_1(c_1, c_2, k, l_1, l_2) = u(c_1) - h(l_1) + \hat{\beta}_1 \delta [u(c_2) - h(l_2) + \delta w(k)],
\]

\[
U_2(c_2, k, l_2) = u(c_2) - h(l_2) + \hat{\beta}_2 \delta w(k),
\]

\[
U_3(c_3) = w(c_3).
\]

Notice that I allow the partially naïve agents to learn from their mistakes and update their beliefs on \( \beta \), so unless the agents are dogmatic, I allow for \( \hat{\beta}_1 \neq \hat{\beta}_2 \). Learning does not occur if the agents start off sophisticated.

The learning process is not explicitly modeled. The only restriction is if \( \hat{\beta}_1 > \beta \), then \( \beta \leq \hat{\beta}_2 < \hat{\beta}_1 \). The agents can be partially naïve for both periods, or be sophisticated at the second period. Therefore, if agents are non-sophisticated in the first period, there are two cases to consider: the partial learning (\( \hat{\beta}_2 \neq \beta \)) case and the full learning (\( \hat{\beta}_2 = \beta \)) case. It is plausible to imagine that agents learn about their biases outside of the model as well. As a result, I will further assume that the learning process is independent of the government policy. For the full learning case, the government fools the agents in the first period and then threaten them in the subsequent period. I will focus on the case with non-sophisticated agents, \( 1 \geq \hat{\beta}_1 > \beta \). Without loss of generality, let \( \delta = 1 \).

For simplicity, let \( \Theta = \{\theta_b, \theta_g\} \) and let the initial distribution be \( 1 > \Pr(\theta_{m,1}) = \pi_m > 0 \), with transition probability \( 1 > \Pr(\theta_{m',2}|\theta_{m,1}) = \pi_{m,m'} > 0 \). The agents only differ in their productivity. They share the same underlying present bias \( \beta \) and initial belief \( \hat{\beta}_1 \). They also share the same learning process, so \( \hat{\beta}_2 \) is the same for all agents. The analysis will begin by discussing the partial learning case.
For the partial learning case, the real allocations are
\[
\{ (c_R^1(\theta_m^1), l_1(\theta_m^1)) ; (c_R^2(\theta_{m'}^2; \theta_m^1), l_2(\theta_{m'}^2; \theta_m^1)) ; k^R(\theta_m^1, \theta_{m'}^2) \}_{m, m' \in \{b, g\}},
\]
and the imaginary allocations are
\[
\{ c_I^1(\theta_m^1), c_I^2(\theta_{m'}^2; \theta_m^1), k^I(\theta_m^1, \theta_{m'}^2) \}_{m, m' \in \{b, g\}},
\]
In the second period, at the reporting stage, any type \((\theta_m^1, \theta_{m'}^2)\) agent faces similar incentive compatibility constraints and fooling constraints as (12), (13), (14), (15), (16) and (17). Since for partial learning \(\hat{\beta}_2 > \beta\), the imaginary allocations can be designed such that the incentive compatibility constraints are non-binding, the government can fool the agents and achieve the desired redistribution without any distortions.

For the first period, in addition to the first period imaginary allocations, the government deceives the partially naïve agents with an imaginary continuation value \(B_I^1(\theta_m^1)\) for any reported first period type \(\theta_m^1\). After the agents supply their labor, the present bias appears and would instead choose \(c_R^1(\theta_m^1)\) and a continuation value of
\[
B(\theta_m^1) = \sum_{\theta_{m'}^2 \in \Theta} \pi_{m, m'} \left[ u(c_I^1(\theta_{m'}^2; \theta_m^1)) - h(l_2(\theta_{m'}^2; \theta_m^1)) + w(k^I(\theta_m^1, \theta_{m'}^2)) \right],
\]
which is the ‘chosen’ continuation value for the \(\theta_m^1\) agent in the first period.

The type \(m\) agent faces the following incentive compatibility constraint in the first period, for any \(\theta_m^1 \in \Theta\),
\[
 u \left( c_I^1(\theta_m^1) \right) - h \left( l_1(\theta_m^1) \right) + B^I(\theta_m^1) \geq u \left( c_I^1(\theta_m^1) \right) - h \left( \frac{\theta_m^1 l_1(\theta_m^1)}{\theta_m^1} \right) + B^I(\theta_m^1),
\]
and the following fooling constraints
\[
 u \left( c_I^1(\theta_m^1) \right) + \hat{\beta}_I B^I(\theta_m^1) \geq u \left( c_I^1(\theta_m^1) \right) + \hat{\beta}_1 B(\theta_m^1),
\]
\[
 u \left( c_R^1(\theta_m^1) \right) + \beta B(\theta_m^1) \geq u \left( c_I^1(\theta_m^1) \right) + \beta B^I(\theta_m^1).
\]
Therefore, with the appropriate imaginary first period consumption and continuation value, the government is able to deceive the agents in the first period and achieve any redistribution.

For the full learning case, since the agents are sophisticated in the second period, the government needs to implement a threat mechanism similar to the one described in Section 4. For the first period, the government can deceive the agents in a similar way as the partial
learning case by using imaginary continuation payoffs and attain perfect insurance across productivity types.

If agents start off sophisticated, then the government implements the threat mechanism in both periods. The threat mechanism in the second period is the same as the one in Section 4 of the paper. In the first period, the government threatens to lower the continuation payoff from $B(\theta^1_m)$ to $B^T(\theta^1_m')$ if an agent with productivity $\theta^1_m$ misreports productivity as $\theta^1_m'$. For the threat to be credible, the government combines $B^T(\cdot)$ with the appropriate $(c^T_1(\theta^1_m), l^T_1(\theta^1_m))$.

References


Ausubel, Lawrence, “Adverse Selection in the Credit Card Market,” *Unpublished*.


