Addiction and Present-Biased Preferences

Ted O'Donoghue
Department of Economics
Cornell University

and

Matthew Rabin
Department of Economics
University of California – Berkeley

May 21, 2001

Abstract

We investigate the role that self-control problems — modeled as time-inconsistent, present-biased preferences — and a person's awareness of those problems might play in leading people to develop and maintain harmful addictions. Present-biased preferences create a tendency to over-consume addictive products, and awareness of future self-control problems can mitigate or exacerbate this over-consumption, depending on the environment. Our central concern is the welfare consequences of this over-consumption. Our analysis suggests that for realistic environments self-control problems are a plausible source of severely harmful addictions only in conjunction with some unawareness of future self-control problems.

Keywords: Addiction, Hyperbolic Discounting, Naivete, Present-Biased Preferences, Self Control, Sophistication, Time Inconsistency.

JEL Classification: A12, B49, C70, D11, D60, D74, D91, E21

Acknowledgments: We are grateful to David Laibson and other participants at the Russell Sage Foundation Conference on Addiction in June 1997, and to seminar participants at Cornell University, U.C. Berkeley, Stanford University, and the 1999 Winter Meetings of the Econometric Society for useful comments, and to Davis Beekman, Kitt Carpenter, Erik Eyster, David Huffman, Ellen Myerson, and Mandar Oak for research assistance. For financial support, we thank the National Science Foundation (Awards SBR-9709485 and SES-0078796), and Rabin thanks the Russell Sage, MacArthur, and Sloan Foundations. This research was started while both authors were visiting the Center for Mathematical Studies in Economics and Management Sciences at Northwestern University. We are very grateful for their hospitality and financial support. This paper is related to our chapter "Addiction and Self Control" in the book *Addiction: Entries and Exits*.

Mail: Ted O'Donoghue / Department of Economics / Cornell University / Ithaca, NY 14853-7601, and Matthew Rabin / Department of Economics / 549 Evans Hall #3880 / University of California, Berkeley / Berkeley, CA 94720-3880. E-mail: edo1@cornell.edu and rabin@econ.berkeley.edu. Web pages: www.people.cornell.edu/pages/edo1/ and http://elsa.berkeley.edu/~rabin/.

1. Introduction

Over the years, researchers from a variety of fields have investigated the consumption of harmful addictive products, such as cigarettes and alcohol, in an attempt to understand why people develop and maintain seemingly destructive addictions. Recently, economists such as Becker and Murphy (1988) have studied rational-choice models of addiction. These models make the natural assumption that people are forward-looking and take into account how current consumption of addictive products will affect their future well-being. But since these models assume in addition that people are "100% rational", they *a priori* rule out a variety of explanations for addictive behavior that many observers consider important. Most non-economists — and we suspect many economists as well — do not view rational-choice models of addiction to be a fully adequate description of why people develop and maintain harmful addictions.

In this paper, we investigate the role that self-control problems — and a person's awareness of those problems — might play in harmful addictions. We delineate the ways in which self-control problems might lead to suboptimal over-consumption of an addictive product. But our main concern is the welfare consequences of this over-consumption, and in particular determining whether self-control problems are a plausible source of severely harmful addictions.

In Section 2, we introduce a model of addiction in which a person decides each period whether to "hit" or "refrain". This binary-choice model is more tractable than previous models, while still incorporating the two crucial characteristics of harmful addictive products found in these previous models. First, harmful addictive products involve *negative internalities*: The more of the product a person has consumed in the past, the lower is his overall well-being now. Second, they involve *habit formation*: The more of the product a person has consumed in the past, the more he desires that product now. The combination of negative internalities and habit formation creates the trap of addiction: As a person consumes more and more of an addictive product, he gets less and less pleasure from its consumption, yet he may continue to consume the product because refraining becomes more and more painful.

We model self-control problems by assuming that people have time-inconsistent *present-biased preferences*, whereby they pursue immediate gratification in ways that do not correspond to their long-run well-being. We apply a simple model of such preferences that was originally proposed by Phelps and Pollak (1968) in the context of intergenerational altruism, and first used by Laibson

(1994,1997) to capture self-control problems within an individual. To examine the role of awareness of future self-control problems, we consider two extreme assumptions about such awareness: *Sophisticates* are fully aware of their future self-control problems, and *naifs* are fully unaware of their future self-control problems. By systematically comparing sophisticates and naifs to people with standard, time-consistent preferences — whom we refer to as *TCs* — we can delineate how predictions depend both on present-biased preferences *per se* and on assumptions about foresight.[1]

In Section 3, we go through an example that illustrates our most basic results. We first show that naifs are always more prone to hit than TCs, reflecting that the direct implication of present-biased preferences is a tendency to over-consume addictive products. Intuitively, the decision whether to consume an addictive product boils down to whether the current desire to consume outweighs the future cost of this consumption, and a preference for immediate gratification makes a person more prone to conclude that hitting is worthwhile. We next show that sophisticates can be more or less prone to hit than naifs, reflecting that awareness can mitigate or exacerbate over-consumption. This ambiguity arises because there are two ways in which awareness can influence current behavior. First, sophisticates are pessimistic about their future behavior, and believe in general that they will hit more often in the future than TCs will (and than naifs think they will). We show in Section 3 that the habit-forming property of addictive goods implies that this *pessimism effect* tends to exacerbate over-consumption due to present-biased preferences. But the pessimism effect can be counteracted by an *incentive effect*: Because sophisticates are worried about improper future over-consumption, they may refrain now in an attempt to induce themselves to resist temptation in the future.

In Section 4, we consider a stationary model of addiction that assumes a person's desire to consume the product depends on past consumption but is otherwise constant over time. While stationarity is unrealistic, it is useful as a base case and to clarify some important intuitions.[2] In this environment, sophisticates are more likely than naifs to develop a harmful addiction, but are also

[1] For other papers on self-control problems and addiction, see Caillaud, Cohen, and Jullien (1996), Carrillo (1999), O'Donoghue and Rabin (1999*b*), and Gruber and Koszegi (2000). Caillaud, Cohen, and Jullien use a different framework to show that if people follow "self-restrained strategies" they might consume in moderation. Carrillo examines sophisticates, and shows that if there is rational uncertainty about negative internalities for addictive products, then sophisticates may abstain so as to avoid learning that the internalities are sufficiently small to justify continued consumption. O'Donoghue and Rabin analyze sophisticates and naifs in a simplified version of the model studied here. Gruber and Koszegi also analyze sophisticates and naifs; their main theoretical conclusions involve optimal cigarette taxation designed to counteract over-consumption due to self-control problems. All these papers, except O'Donoghue and Rabin's, assume that consumption is a continuous choice, and for simplicity limit attention to stationary environments. See also Elster (1999).

[2] Stationarity is assumed in all rational-addiction models with which we are familiar, and also in Caillaud, Cohen, and Jullien (1996), Carrillo (1999), and Gruber and Koszegi (2000).

more likely than naifs to quit an established addiction. These results reflect the interplay between the pessimism and incentive effects in the stationary environment, and in particular how the incentive effect is stronger the more addicted a person is. We then ask whether self-control problems represent a plausible source of *severe* harm, and identify two potential sources of severe harm in the stationary model. First, to the extent that people are sophisticated, they may suffer severe harm due to feelings of inevitability. Even when a person would prefer non-addiction, if he thinks he'll get addicted in the future no matter what he does today, he may conclude that he might as well start consuming today. Second, to the extent that people are naive, they may suffer severe harm from procrastination in quitting. Even when quitting is well worth it, if the person prefers quitting in the near future rather than now, he may repeatedly delay quitting.

In Section 5, we relax the unrealistic assumption of stationarity, and explore a "youthful" model in which for any given addiction level the temptation to hit is larger earlier in life than later in life. We use this model to show that the stationary model yields overly pessimistic predictions with regard to sophisticates and overly optimistic predictions with regard to naifs. In the stationary model, sophisticated self-control problems are problematic when they cause a person to feel that addiction is inevitable; in the youthful model, inevitability is less likely — in particular, there is no inevitability under the plausible assumption that the person eventually matures to a point where he would have no desire to consume if he were unaddicted at that time. In the stationary model, naive self-control problems are problematic when they cause a person to procrastinate quitting an established addiction; in the youthful model, large initial temptations provide the catalyst for establishing addictions which naifs never quit.

We also use the youthful model to explore the role of temporary temptations in developing harmful addictions. While Becker and Murphy (1988) show how it can be optimal for a person to *maintain* a severely harmful addiction, their steady-state model provides no formal analysis of why the person would choose to *develop* this harmful addiction in the first place.[3] In their informal discussion, Becker and Murphy suggest events such as youth, divorce, and the death of a loved one as possible sources of harmful addictions. The youthful model permits us to directly investigate this

[3] This shortcoming has been recognized in the rational-addiction literature, and fixes have been proposed. Orphanides and Zervos (1995) and Wang (1997) posit that people might develop harmful addictions due to rational uncertainty about the addictiveness of a product. Suranovic, Goldfarb, and Leonard (1999) and Goldbaum (2000) posit that people might get addicted while young and later quit because the detrimental effects of consumption occur mainly at the end of a person's life. We feel there is some truth to both these stories, but we also feel that they are complements — and not substitutes — to our approach.

hypothesis, and in particular to ask whether a person would indulge his short-term desire despite its long-term consequences. We show that while such events can clearly cause an addiction for all three types, such an addiction can be severely harmful only for naifs.

Our analysis in Sections 4 and 5 assumes that prices are held constant; in Section 6 we explore the effects of price on consumption. Although consumption is a discrete choice in our model, and therefore our analysis of price comparative statics is necessarily crude, we are able to capture some important intuitions for TCs and naifs. While the qualitative effects of price changes are the same for TCs and naifs, our model predicts different quantitative effects. In particular, because naifs underestimate their own future consumption, the effects of future prices on current consumption are much smaller for naifs than for TCs. This intuition might provide an explanation for the puzzle in the empirical literature on rational addiction that temporary price changes and permanent price changes have similar effects on consumption. Under the maintained hypothesis of time consistency, this empirical result implies that people have absurd discount rates. But our model suggests that this empirical result might be consistent with a reasonable long-term discount rate combined with a small self-control problem about which the person is naive.

Finally, we conclude in Section 7 with a discussion of some general lessons to take away from our analysis, with an emphasis on why we feel our model of addiction and present-biased preferences is an improvement on rational-choice models of addiction.

2. The Model

The crucial feature of addictive products is that past consumption affects current well-being. Becker and Murphy (1988) provide a model of instantaneous utility functions that captures this feature.[4] In this paper, we introduce a simplified version of their model: Whereas most models of addiction follow Becker and Murphy (1988) in assuming consumption is a continuous choice, we model consumption as a binary choice. Our model maintains the key features of Becker and Murphy's model, and our main conclusions are driven by these features. But by assuming a less realistic binary choice, our model is significantly more tractable, allowing us to solve for optimal behavior rather than merely steady-state behavior, and permitting analysis of a richer array of environments. Most importantly, we are able to directly analyze the role of non-stationarities in the temptation to

[4] For earlier work on habit formation using a similar formulation, see Pollak (1970) and Ryder and Heal (1973).

consume, which seem likely to play an important role in why people develop harmful addictions.

We consider a discrete-time model with periods $1, ..., T$, where we consider both $T < \infty$ and $T = \infty$. Each period, a person can either take a "hit", in which case his consumption $a_t = 1$, or "refrain", in which case $a_t = 0$. In a given period, the person decides only whether to hit now, and has no way to commit to future behavior. For most of our analysis, we assume that the addictive product is free, which helps highlight the fact that people may avoid addictive products not because of their purchase price *per se*, but rather because of their detrimental long-run consequences. We explore the role of prices for consumption in Section 6.

Let k_t be the person's *addiction level* in period t, which captures all effects of past consumption for period-t instantaneous utility. We assume k_t evolves according to the equation $k_t = \gamma k_{t-1} + a_{t-1}$, where $\gamma \in [0, 1)$ is a parameter indicating the rate at which an addiction decays. When $\gamma = 0$, refraining for a single period gets the person completely unaddicted. For γ close to 1, refraining reduces the person's addiction level very little. The appropriate γ depends on both the nature of the addictive product being examined, as well as on the time scale of each "period", be it a day, a year, or an epoch of one's life. This formulation implies a maximum addiction level: If the person hits every period, his addiction level converges to $k^{\max} \equiv \sum_{t=1}^{\infty} \gamma^{t-1} = \frac{1}{1-\gamma}$.[5]

We assume the person's instantaneous utility function in period t is

$$u_t(a_t, k_t) \equiv \begin{cases} x_t + f(k_t) & \text{if } a_t = 1 \\ y_t + g(k_t) & \text{if } a_t = 0. \end{cases}$$

Without loss of generality, we set $f(0) = g(0) = 0$, and we often drop the subscript t from k_t and a_t when there is no danger of confusion. This formulation allows for the instantaneous utility function to be constant across time or to vary.

The *temptation to hit* in period t is $h_t(k) \equiv u_t(1, k) - u_t(0, k) = [x_t - y_t] + [f(k) - g(k)]$, which is the person's instantaneous marginal utility from hitting. The temptation to hit consists of two components: an exogenous component $x_t - y_t \equiv \bar{x}_t$ that is independent of past consumption, and an endogenous component $f(k) - g(k)$ that depends on past consumption.

Our analysis hinges on two characteristics of addictive products. First, they generate negative internalities: The more the person has consumed in the past, the smaller is his current well-being. Negative internalities include health, job, and personal problems caused by past consumption. Neg-

[5] The parameter γ corresponds to $(1 - \delta)$ in Becker and Murphy (1988). This formulation is potentially restrictive in that it combines into a single parameter the rate at which a person becomes addicted when hitting and the rate at which a person becomes unaddicted when refraining.

ative internalities also include "tolerance" — the loss in enjoyment of an addictive substance due to regular consumption.[6] Formally:

Definition 1. A product has **negative internalities** if for all k, $f'(k) < 0$ and $g'(k) < 0$.

In addition to generating negative internalities, addictive products are habit-forming: The more of the product the person has consumed in the past, the more he will be tempted to consume now.[7] Formally:

Definition 2. A product is **habit-forming** if for all t and k, $h'_t(k) = f'(k) - g'(k) > 0$.

Although negative internalities are incorporated into both f and g, we often refer to $f(k) \leq 0$ as the internality cost of past consumption, and $g(k) - f(k) \leq 0$ as the additional cost of past consumption due to habit formation. A person incurs the internality cost $f(k)$ no matter what he does in period t; he incurs the additional cost $g(k) - f(k)$ only if he refrains in period t. Hence, habit formation implies that the cost of past consumption is larger when the person refrains as opposed to hit. This feature of addictive products will play an important role in our analysis.

Besides assuming negative internalities and habit formation, we assume that f and g are weakly convex in k: $f''(k) \geq 0$ and $g''(k) \geq 0$.[8] The more addicted the person becomes, the less a given increase in k hurts his instantaneous utility, and therefore the less harm hitting does to future utility. In fact, we shall often assume that the instantaneous utility function takes the following linear form:

$$u_t(a, k) \equiv \begin{cases} x_t - \rho k & \text{if } a = 1 \\ y_t - (\rho + \sigma)k & \text{if } a = 0. \end{cases}$$

In this formulation, the parameter $\rho > 0$ represents the internality cost, and the parameter $\sigma > 0$ represents the additional cost due to habit formation.

The trade-off between the temptation to hit and its future costs is the crux of the choice to become addicted. How people weigh this trade-off depends on their intertemporal preferences. The

[6] We borrow the term "internalities" from Herrnstein, Loewenstein, Prelec, and Vaughn (1993), who define an internality to be a "within-person externality". Of course, some products generate positive internalities, in particular learning and other "investment goods" that have long-term benefits. But harmful addictive products are generally thought to generate negative internalities, and that is the case we focus on in this paper.

[7] Internalities and habit formation are not inherently tied together; eating cheesecake may generate negative internalities, but is not necessarily habit-forming.

[8] Most results hold even if f and g are a little concave, and some do not rely at all on them being convex. Note that we make no assumption about whether the endogenous temptation $f(k) - g(k)$ is convex or concave.

standard economics model assumes that intertemporal preferences are *time-consistent*: A person's relative preference for well-being at an earlier date over a later date is the same no matter when she is asked. But there is a mass of evidence that intertemporal preferences take on a specific form of *time inconsistency*: A person's relative preference for well-being at an earlier date over a later date gets stronger as the earlier date gets closer.[9] In other words, people have self-control problems caused by a tendency to pursue immediate gratification in a way that their "long-run selves" do not appreciate.

In this paper, we apply a simple form of such *present-biased preferences*, using a model originally developed by Phelps and Pollak (1968) in the context of intergenerational altruism, and first used by Laibson (1994,1997) to capture self-control problems within an individual.[10] Let u_t be the instantaneous utility the person gets in period t. Then his intertemporal preferences from the perspective of time t can be represented by the following intertemporal utility function:

$$U^t(u_t, u_{t+1}, ..., u_T) \equiv u_t \ + \ \beta \sum_{\tau=t+1}^{T} \delta^{\tau-t} u_\tau.$$

The parameter δ represents "time-consistent" discounting, while the parameter β represents the "present bias". For $\beta = 1$ these preferences reduce to (the discrete version of) exponential discounting, whereas for $\beta < 1$ these preferences parsimoniously capture the time-inconsistent preference for immediate gratification.[11]

To analyze the role of awareness of future self-control problems, we consider two types of people representing extreme assumptions about such awareness: *Sophisticates* are fully aware of their future self-control problems; and *naifs* are fully unaware of their future self-control problems.[12] We

[9] See, for instance, Ainslie (1975, 1991, 1992), Ainslie and Haslam (1992*a*, 1992*b*), Loewenstein and Prelec (1992), Thaler (1991), and Thaler and Loewenstein (1992). While the rubric of "hyperbolic discounting" is often used to describe such preferences, we use the term "present-biased preferences" to reflect the qualitative feature of the time inconsistency that is more general, and more generally supported by empirical evidence, than the specific hyperbolic functional form.

[10] This model has since been used by Laibson (1996), Harris and Laibson (forthcoming), O'Donoghue and Rabin (1999*a*, 1999*b*, 1999*c*, 2001), Fischer (1997), Carrillo (1999), Carrillo and Mariotti (2000), Gruber and Koszegi (2000), and others.

[11] We often refer to the time-consistent discount factor δ *not* as a preference parameter, but rather as a "relevance" parameter, interpreting $1 - \delta$ as the probability of dying between periods t and $t + 1$. But none of our results depend on this interpretation of δ.

[12] These assumptions (and the labels) were originally laid out by Strotz (1956) and Pollak (1968). While there is limited evidence, people clearly exhibit elements of both sophistication and naivete. Most papers studying time-inconsistent preferences assume sophistication (e.g., Laibson (1994, 1996, 1997), Harris and Laibson (forthcoming), Fischer (1997), Carrillo (1999), Carrillo and Mariotti (2000)). Akerlof (1991), O'Donoghue and Rabin (1999*a*, 1999*b*, 1999*c*), and Gruber and Koszegi (2000) also consider naive beliefs. O'Donoghue and Rabin (2001) formalize and analyze a case of partial naivete in between these two extremes.

also analyze standard time-consistent agents, whom we refer to as *TCs*. It is a useful benchmark to understand how TCs would behave, and moreover understanding the behavior of TCs provides a useful analytical tool for understanding the behavior of naifs. But most importantly, the behavior of TCs represents how sophisticates and naifs would like to behave if asked from some prior perspective (before period 1). We make use of this last point in our welfare analysis.

To analyze the behavior of these three types of people, we assume people follow *perception-perfect strategies* (O'Donoghue and Rabin, 1999*a*). In words, a person chooses to hit in period t if and only if hitting in period t is optimal given his period-t preferences and his period-t beliefs about how he will behave in the future. In order to formally describe both behavior and beliefs, we define a *strategy* as a function $\alpha : [0, k^{\max}] \times \{1, 2, ..., T\} \rightarrow \{0, 1\}$, where strategy α prescribes action $\alpha(k, t)$ in period t when the addiction level is k.

Define $U_t(k_t, \alpha)$ to be the person's period-t long-run utility from following strategy α given period-t addiction level k_t. "Long-run utility" represents the person's intertemporal preferences from a prior perspective that is temporally removed from the current desire for immediate gratification — that is, the person's intertemporal preferences when $\beta = 1$. A useful way to write $U_t(k_t, \alpha)$ is to break it down into the immediate instantaneous utility and the intertemporal utility beginning next period:

$$U_t(k_t, \alpha) = \begin{cases} [x_t + f(k_t)] + \delta U_{t+1}(\gamma k_t + 1, \alpha) & \text{if } \alpha(k_t, t) = 1 \\ [y_t + g(k_t)] + \delta U_{t+1}(\gamma k_t, \alpha) & \text{if } \alpha(k_t, t) = 0. \end{cases}$$

Consider a person in period t whose current addiction level is k, and suppose this person perceives that he will follow strategy α^p beginning in period $t + 1$. This person believes that if he hits this period then his intertemporal utility beginning next period will be $U_{t+1}(\gamma k+1, \alpha^p)$, and that if he refrains this period then his intertemporal utility beginning next period will be $U_{t+1}(\gamma k, \alpha^p)$. Hence, he perceives the (undiscounted) future cost from hitting to be $U_{t+1}(\gamma k, \alpha^p) - U_{t+1}(\gamma k + 1, \alpha^p)$. He would then hit in period t if and only if the current temptation to hit $h_t(k)$ is larger than the (discounted) future cost from hitting. For simplicity, we assume a person hits when indifferent.

Given this framework, we can formally define perception-perfect strategies for the three types of people. Because TCs correctly predict their future behavior, and because TCs discount the future

cost of hitting by δ, we define perception-perfect strategies for TCs as:

Definition 3. A **perception-perfect strategy for TCs** is a strategy α^{tc} that satisfies for all $k \geq 0$ and for all t, $\alpha^{tc}(k,t) = 1$ if and only if $h_t(k) \geq \delta \left[U_{t+1}\left(\gamma k, \alpha^{tc}\right) - U_{t+1}\left(\gamma k + 1, \alpha^{tc}\right)\right]$.

At any point in time, naifs believe they will behave like TCs beginning next period — that is, in any period naifs perceive that they will follow strategy α^{tc} beginning next period. Because naifs discount the future cost from hitting by $\beta\delta$, we define perception-perfect strategies for naifs as:

Definition 4. A **perception-perfect strategy for naifs** is a strategy α^{n} that satisfies for all $k \geq 0$ and for all t, $\alpha^{n}(k,t) = 1$ if and only if $h_t(k) \geq \beta\delta \left[U_{t+1}\left(\gamma k, \alpha^{tc}\right) - U_{t+1}\left(\gamma k + 1, \alpha^{tc}\right)\right]$.

Sophisticates, like TCs, predict exactly how they will behave in the future. But sophisticates, like naifs, discount the future cost from hitting by $\beta\delta$. Hence, we define perception-perfect strategies for sophisticates as:

Definition 5. A **perception-perfect strategy for sophisticates** is a strategy α^{s} that satisfies for all $k \geq 0$ and for all t, $\alpha^{s}(k,t) = 1$ if and only if $h_t(k) \geq \beta\delta \left[U_{t+1}\left(\gamma k, \alpha^{s}\right) - U_{t+1}\left(\gamma k + 1, \alpha^{s}\right)\right]$.

For TCs and naifs, this solution concept is equivalent to them formulating an optimal consumption path in each period and choosing the current action that is part of that consumption path.[13] TCs always stick to the consumption path chosen in the first period, whereas naifs often revise their chosen consumption paths as their preferences change from period to period. For sophisticates, in contrast, this solution concept implies that they are in a sense playing a game against their future selves. Hence, their behavior partly reflects "strategic" reactions to bad behavior by future selves that they cannot directly control, and partly reflects attempts to induce good behavior from future selves.[14]

Because our goal is to analyze the role of self-control problems in generating harmful addictions, our analysis ignores a variety of other "errors" that might be important for addiction. We assume

[13] We assume throughout that an optimal consumption path exists.

[14] Conspicuously absent from our model is the use of external commitment devices. Alcoholics sophisticated about their self-control problems may, for instance, choose to check themselves into the Betty Ford Clinic. Note that the existence of external commitment devices would not affect the behavior of naifs (or TCs) since they believe they will behave themselves in the future and therefore see no need for commitment devices.

throughout, for instance, that people correctly predict how the temptation to consume evolves over time, and that people correctly predict how current consumption affects future instantaneous utility functions.[15] We leave the analysis of other errors, and how self-control problems might interact with those errors, for future research.

3. An Example and Some Basic Results

In this section we work through an example that illustrates some important intuitions, and in the process we derive some basic results that hold for any instantaneous utilities satisfying the assumptions outlined in Section 2.

Example 1

Suppose $T = \infty$ and $\gamma = 0$, which implies $k_t \in \{0, 1\}$ for all t — in each period the person is either "unhooked" or "hooked". Suppose further that the person has a linear instantaneous utility function with parameters ρ and σ. Finally, suppose $y_t = 0$ for all t and $x_t = x_o$ for all $t \in \{2, 3, ...\}$. How does an unhooked person behave in period 1 as a function of x_1?

A person hits when the temptation to hit is larger than the perceived future cost of that hit. In period 1, the temptation to hit for an unhooked person is x_1. The future cost of hitting depends on perceived future behavior. Suppose that optimal behavior beginning in period 2 is to refrain in all future periods whether unhooked or hooked at that time.[16] Because TCs (correctly) and naifs (possibly incorrectly) believe that they will behave in this way, they both perceive that hitting will lead to continuation utility $U_2(1, \alpha^{tc}) = -(\rho+\sigma)+\frac{\delta}{1-\delta}0$ and that refraining will lead to continuation utility $U_2(0, \alpha^{tc}) = 0 + \frac{\delta}{1-\delta}0$. Hence, they both perceive the future cost of hitting in period 1 to be $U_2(0, \alpha^{tc}) - U_2(1, \alpha^{tc}) = \rho + \sigma$.

Applying Definitions 3 and 4, TCs hit in period 1 if and only if $x_1 \geq \delta(\rho + \sigma)$, and naifs hit in period 1 if and only if $x_1 \geq \beta\delta(\rho + \sigma)$. Hence, naifs are more prone to hit in period 1 than TCs,

[15] Orphanides and Zervos (1995) and Wang (1997) explore how fully rational people might become addicted because they have incomplete information about the addictiveness of products. Loewenstein, O'Donoghue, and Rabin (2000) study a general form of misprediction of future preferences which when applied to addiction predicts that even a time-consistent person might get harmfully addicted because he mispredicts the addictiveness of products.

[16] As our analysis in Section 4 will reveal, this holds when $-(\rho + \sigma) > (x_o - \rho)/(1 - \delta)$.

reflecting that the direct implication of present-biased preferences is a tendency to over-consume harmful addictive products. This outcome is a straightforward implication of the fact that TCs and naifs perceive the same future implications of hitting, combined with the fact that naifs have a greater taste for immediate gratification. Clearly this conclusion is quite general: Part 3 of Lemma 1 establishes that for any instantaneous utilities satisfying our assumptions in Section 2, in any situation naifs are more likely to hit than TCs. As a preliminary step that will prove quite useful in our later analysis, Lemma 1 also establishes that both TCs and naifs follow cutoff strategies where in each period the person hits if and only if his addiction level is larger than some critical level. This result is driven by the non-concavity of the instantaneous utility function with respect to the addiction level k, which implies that the future cost of hitting is non-increasing in k.[17]

Lemma 1. For any instantaneous utilities and for any T:

(1) There is a unique perception-perfect strategy for TCs, α^{tc}, and for all t there exists $\bar{k}_t^{tc}$ such that $\alpha^{tc}(k,t) = 1$ if and only if $k \geq \bar{k}_t^{tc}$,

(2) There is a unique perception-perfect strategy for naifs, α^{n}, and for all t there exists $\bar{k}_t^{n}$ such that $\alpha^{n}(k,t) = 1$ if and only if $k \geq \bar{k}_t^{n}$, and

(3) $\alpha^{tc}(k,t) \leq \alpha^{n}(k,t)$ for all k and t, or equivalently $\bar{k}_t^{tc} \geq \bar{k}_t^{n}$ for all t.

We next investigate how awareness of future self-control problems affects this over-consumption. In Example 1, naifs in period 1 optimistically believe they will behave themselves in the future — and refrain forever after — whereas sophisticates correctly predict that they may misbehave. Let's first suppose that sophisticates will in fact hit forever after regardless of whether they enter period 2 unhooked or hooked.[18] Because they correctly predict this future behavior, sophisticates perceive that hitting will lead to continuation utility $U_2(1,\alpha^s) = (x_o - \rho) + \frac{\delta}{1-\delta}(x_o - \rho)$ and that refraining will lead to continuation utility $U_2(0,\alpha^s) = x_o + \frac{\delta}{1-\delta}(x_o - \rho)$. Hence, sophisticates perceive the future cost of hitting in period 1 to be $U_2(0,\alpha^s) - U_2(1,\alpha^s) = \rho$.

Applying Definition 5, sophisticates hit in period 1 if and only if $x_1 \geq \beta\delta\rho$. Given our earlier conclusion that naifs perceive the future cost of hitting to be $\rho + \sigma$, and hence hit in period 1 if and only if $x_1 \geq \beta\delta(\rho + \sigma)$, in this case sophisticates are *more* prone to hit in period 1 than naifs. This outcome reflects the implications of pessimism in the realm of addiction: Because the

[17] All proofs are in the Appendix.

[18] As our later analysis will reveal, this holds when $x_o \geq \beta\delta\rho$.

habit-forming property of addictive products implies that a current hit has a larger future cost the more one expects to refrain in the future, pessimism about future behavior makes a person more prone to succumb to the temptation to hit, and therefore tends to exacerbate over-consumption. We refer to this logic as the *pessimism effect*, and Lemma 2 establishes that this logic holds for any instantaneous utilities satisfying our assumptions in Section 2.

Lemma 2. Suppose that for both $k_{t+1} = \gamma k_t$ and $k_{t+1} = \gamma k_t + 1$, strategy α induces consumption path $(a_{t+1}, a_{t+2}, ..., a_T)$ and strategy α' induces consumption path $(a'_{t+1}, a'_{t+2}, ..., a'_T)$. If $a_\tau \geq a'_\tau$ for all $\tau \geq t+1$, then $U_{t+1}(\gamma k_t, \alpha) - U_{t+1}(\gamma k_t + 1, \alpha) \leq U_{t+1}(\gamma k_t, \alpha') - U_{t+1}(\gamma k_t + 1, \alpha')$.

Lemma 2 states that if for both α and α' the future consumption path is independent of current consumption, and if α involves unambiguously more future consumption than α', then a current hit causes less future harm under α. This result plays an important role in the implications of sophistication. If future behavior does not depend on current behavior, so that the implications of sophistication derive solely from different perceptions of how much they will consume in the future, naifs are less likely to consume than sophisticates.

There is more to sophistication than simple pessimism, however, because current behavior might influence future behavior. Let us again return to Example 1, but now suppose that sophisticates will hit forever after if they are hooked in period 2 but will refrain forever after if they are unhooked in period 2.[19] Sophisticates now perceive that hitting will lead to continuation utility $U_2(1, \alpha^s) = (x_o - \rho) + \frac{\delta}{1-\delta}(x_o - \rho)$ and that refraining will lead to continuation utility $U_2(0, \alpha^s) = 0 + \frac{\delta}{1-\delta}0$. Hence, they perceive the future cost of hitting in period 1 to be $U_2(0, \alpha^s) - U_2(1, \alpha^s) = \frac{\rho - x_o}{1-\delta}$, and therefore hit in period 1 if and only if $x_1 \geq \beta\delta\left(\frac{\rho - x_o}{1-\delta}\right)$. Given our earlier conclusion that naifs hit in period 1 if and only if $x_1 \geq \beta\delta(\rho + \sigma)$, and given our presumption that optimal behavior is to refrain forever after even if hooked, which implies $-(\rho+\sigma) > \frac{x_o - \rho}{1-\delta}$, we conclude that sophisticates are *less* prone to hit in period 1 than naifs.

In Example 1, sophisticates might refrain in period 1 while naifs hit if sophisticates are refraining in an attempt to induce future restraint — or equivalently in an attempt to prevent future misbehavior. We refer to this second effect of sophistication as the *incentive effect*: Because sophisticates are worried about improper future over-consumption, they may refrain now in an attempt to induce

[19] As our later analysis will reveal, this holds when $\beta\delta\rho > x_o \geq \rho - \left(\frac{1-\delta}{1-\delta+\beta\delta}\right)\sigma$.

themselves to resist temptation in the future. In the realm of addiction, the incentive effect means that sophistication can mitigate over-consumption due to present-biased preferences. Hence, there is a tension between the pessimism and incentive effects that determines whether sophisticates are more or less prone to consume than naifs. In the next two sections, we examine how this tension plays out in some different environments.[20]

4. Stationary Preferences

Preferences are stationary when a person's instantaneous utility function $u_t(a,k)$ depends on his current addiction level k but not on the specific period t. Formally:

Stationary Preferences:

$$\text{For all } t, \qquad u_t(a,k) \equiv \begin{cases} x_o + f(k) & \text{if } a = 1 \\ y_o + g(k) & \text{if } a = 0. \end{cases}$$

Stationary preferences are not particularly realistic. Such preferences mean, for instance, that the first hit of a cigarette or cocaine yields the same pleasure to a 20-year old as it does to a 60-year old. On both social and physiological grounds we are skeptical of this assumption. But stationary preferences are useful as a base case and to clarify some important intuitions.

Our analysis of stationary preferences assumes an infinite horizon, in part for expositional ease, and in part because this assumption is closer in spirit to the rational-choice models of addiction. In addition, we often analyze how a person behaves starting from an initial addiction level $k_1 > 0$, which can be naturally interpreted as reflecting the net effects of unmodeled past consumption. Indeed, our analysis here of the $k_1 > 0$ case is a useful building block for our analysis in the next section, where we look at a stationary model preceded by a youthful period of larger exogenous temptations. Finally, many of our results in this section will be stated in terms of $\bar{x}_o \equiv x_o - y_o$.

Lemma 3 establishes that with stationary preferences and an infinite horizon, the cutoff for TCs and the cutoff for naifs are both stationary.

Lemma 3. Under stationary preferences and $T = \infty$, there exists $\bar{k}^{tc}$ such that for all t, $\alpha^{tc}(k,t) = 1$ if and only if $k \geq \bar{k}^{tc}$; and there exists $\bar{k}^{n}$ such that for all t, $\alpha^{n}(k,t) = 1$ if and only if $k \geq \bar{k}^{n}$.

[20] The pessimism and incentive effects are first discussed in O'Donoghue and Rabin (1999*b*). These effects represent a decomposition of the "sophistication effect" identified in O'Donoghue and Rabin (1999*a*).

Intuitively, TCs and naifs choose optimal consumption paths, and with stationary preferences and an infinite horizon the optimal consumption paths are independent of the current period. An immediate implication of Lemma 3 is that for any initial addiction level both TCs and naifs either never hit or hit always.

For sophisticates, unlike for TCs and naifs, there can be multiple perception-perfect strategies for an infinite horizon. We restrict attention to perception-perfect strategies for the infinite horizon that correspond to the unique finite-horizon perception-perfect strategy as the horizon becomes long.[21] Lemma 4 characterizes the behavior of sophisticates under this restriction:

Lemma 4. Under stationary instantaneous utilities and $T = \infty$:

(1) If $\bar{x}_o \geq \beta\delta\Delta^H$ then $\alpha^s(k,t) = 1$ for all k and t; and

(2) If $\bar{x}_o < \beta\delta\Delta^H$ then there exists $k' > 0$ such that for all $k < k'$ $\alpha^s(k,t) = 0$ for all t, where

$$\Delta^H \equiv \sum_{n=1}^{\infty} \delta^{n-1} \left[f\left(\sum_{m=1}^{n-1} \gamma^{m-1} \right) - f\left(\sum_{m=1}^{n} \gamma^{m-1} \right) \right].$$

The value Δ^H is the future cost from hitting for an unaddicted person who has the most pessimistic beliefs possible: He believes he will hit forever ever after no matter what he does now. Lemma 4 establishes that a crucial question for sophisticates is how they would behave when unaddicted given such extremely pessimistic beliefs. Intuitively, if there is a finite horizon, a sophisticate facing exogenous temptation $\bar{x}_o > 0$ recognizes that he will hit in the final moments of his life.[22] Hence, if he refrains at all, it must be that there is some moment far enough from the end of his life where he prefers to refrain despite pessimistically believing he will hit for the remainder of his life, which holds if and only if $\bar{x}_o < \beta\delta\Delta^H$. The condition $\bar{x}_o \geq \beta\delta\Delta^H$ can be interpreted as a kind of "inevitability condition": If it holds, sophisticates perceive that addiction is inevitable in the sense that no matter what they do today their future selves will hit forever after. An immediate

[21] For both TCs and naifs, the unique infinite-horizon perception-perfect strategy corresponds to the unique finite-horizon perception-perfect strategy as the horizon becomes long. For sophisticates, this restriction rules out infinite-horizon perception-perfect strategies where a person refrains because of a belief that hitting will lead to bad continuation utility beyond the change in incentives, analogous to folk-theorem type equilibria in infinitely-repeated games. The reader should not be overly worried about this restriction because it biases sophisticates towards "bad behavior" — it rules out strategies whereby sophisticates behave themselves due to this mentality — and yet we shall conclude that in realistic environments sophisticated self-control problems are not a plausible source of severe addictions.

[22] While the discussion in the text uses the assumption $\bar{x}_o > 0$ to avoid some technical details, the formal results do not rely on this assumption, as $\bar{x}_o < 0$ merely implies we are in the case $\bar{x}_o < \beta\delta\Delta^H$.

implication of Lemma 4 is that an unaddicted sophisticate either never hits or hits always.

It will prove useful in deriving a person's *actual* behavior path to consider the person's *desired* behavior path. Lemma 3 implies that for any situation the desired behavior path for TCs is either hitting always or never hitting. While the desired behavior path for a person with present-biased preferences also might be never hitting or hitting always, a third possibility arises: The person might want to hit now and never again. Let $k^*(\beta)$ denote the addiction level such that a person prefers hitting always to never hitting if and only if $k \geq k^*(\beta)$, and let $\tilde{k}(\beta)$ denote the addiction level such that a person prefers hitting once to never hitting if and only if $k \geq \tilde{k}(\beta)$. Lemma 5 uses these values to describe behavior for the three types.[23]

Lemma 5. Under stationary instantaneous utilities and $T = \infty$, for all t:

(1) $\alpha^{tc}(k,t) = 1$ if and only if $k \geq k^*(1)$;

(2) $\alpha^{n}(k,t) = 1$ if and only if $k \geq \min\{k^*(\beta), \tilde{k}(\beta)\}$; and

(3) If $\bar{x}_o \geq \beta\delta\Delta^H$, then $\alpha^{s}(k,t) = 1$ for all k; if $\bar{x}_o < \beta\delta\Delta^H$ and $\gamma\tilde{k}(\beta) + 1 \geq k^*(\beta)$, then $\alpha^{s}(k,t) = 1$ if and only if $k \geq k^*(\beta)$; and if $\bar{x}_o < \beta\delta\Delta^H$ and $\gamma\tilde{k}(\beta) + 1 < k^*(\beta)$, then $\alpha^{s}(k,t) = 0$ if $k < \tilde{k}(\beta)$ and $\alpha^{s}(k,t) = 1$ if $k \geq k^*(\beta)$.

Part 1 characterizes the actual behavior of TCs. Because for TCs actual behavior is identical to desired behavior, TCs hit if and only if they prefer hitting always to never hitting, which holds if and only if their current addiction level k is larger than $k^*(1)$. Part 2 characterizes the actual behavior of naifs. Because naifs attempt to follow their desired behavior path, they hit if and only if they prefer either hitting always or hitting once to never hitting, which holds if and only if their current addiction level k is larger than either $k^*(\beta)$ or $\tilde{k}(\beta)$.

Part 3 characterizes the actual behavior of sophisticates. If $\bar{x}_o \geq \beta\delta\Delta^H$ then, as established by Lemma 4, sophisticates hit no matter what. If $\bar{x}_o < \beta\delta\Delta^H$, then sophisticates refrain whenever their desired behavior is never hitting, which holds if $k < \min\{k^*(\beta), \tilde{k}(\beta)\}$, and sophisticates hit whenever hitting always is preferred to never hitting, which holds if $k \geq k^*(\beta)$. The remaining question is how do sophisticates behave if $k \in [\tilde{k}(\beta), k^*(\beta))$. In this case, sophisticates would like to hit once, but if $\gamma\tilde{k}(\beta) + 1 \geq k^*(\beta)$ then a single hit would increase their addiction level to the

[23] The proof of Lemma 5 provides equations defining $k^*(\beta)$ and $\tilde{k}(\beta)$. Because for TCs hitting once is never desired, clearly $k^*(1) \leq \tilde{k}(1)$. For $\beta < 1$, $k^*(\beta)$ and $\tilde{k}(\beta)$ are not rankable; but hitting once can be desired only if $\tilde{k}(\beta) < k^*(\beta)$.

point where they would hit forever after. Hence, hitting once (or any finite number) is not feasible, and so sophisticates refrain for any $k \in [\tilde{k}(\beta), k^*(\beta)]$. If instead $\gamma\tilde{k}(\beta) + 1 < k^*(\beta)$, there can be situations in which hitting for a finite number of periods is feasible, in which case sophisticates' behavior can be quite complicated for $k \in [\tilde{k}(\beta), k^*(\beta))$. In fact, because of these complications sophisticates need not follow a stationary strategy or a cutoff strategy.

Lemma 5 permits a simple comparison of the behavior of TCs, naifs, and sophisticates, which we present as Proposition 1:

Proposition 1. Under stationary instantaneous utilities and $T = \infty$:

(1) If $\bar{x}_o \geq \beta\delta\Delta^H$, then $\alpha^{tc}(k,t) \leq \alpha^n(k,t) \leq \alpha^s(k,t)$ for all k and t; and

(2) If $\bar{x}_o < \beta\delta\Delta^H$, then $\alpha^{tc}(k,t) \leq \alpha^s(k,t) \leq \alpha^n(k,t)$ for all k and t.

Proposition 1 establishes that in the stationary model whether sophistication makes a person more or less likely to consume an addictive product depends crucially on whether the inevitability condition $\bar{x}_o \geq \beta\delta\Delta^H$ holds. This conclusion reflects the interplay between the pessimism and incentive effects in the stationary model. The pessimism effect is always at work in inducing sophisticates to consume more than naifs. The crucial question therefore is under what conditions is the incentive effect operative, leading sophisticates to refrain in order to induce good behavior in the future. Refraining now can induce future restraint only if persistent restraint puts the person in a situation where he would refrain even in the absence of the incentive effect. With stationary instantaneous utilities, such situations are possible if and only if $\bar{x}_o < \beta\delta\Delta^H$. Proposition 1 establishes that whenever $\bar{x}_o < \beta\delta\Delta^H$, the incentive effect is operative and sophisticates refrain whenever naifs refrain.

Whereas Proposition 1 describes how the implications of sophistication depend on the inevitability condition, Proposition 2 describes how the implications of sophistication depend on the initial addiction level.

Proposition 2. Under stationary instantaneous utilities and $T = \infty$, for any β, δ, γ, $f(\cdot)$, and $g(\cdot)$, there exists $\bar{k} \in (0, k^{\max})$ such that

(1) If $k_1 \leq \bar{k}$, then for any $\bar{x}_o$ where naifs hit always, sophisticates hit always; and

(2) If $k_1 \geq \bar{k}$, then for any $\bar{x}_o$ where naifs never hit, sophisticates never hit.

Proposition 2 establishes that for sufficiently unaddicted people, sophisticates are more likely to hit always than naifs, whereas for sufficiently addicted people, naifs are more likely to hit always than sophisticates. For the case of continuous consumption (and with additional assumptions about functional forms), Gruber and Koszegi (2000) find a similar result. These results once more reflect the interplay between the pessimism and incentive effects. As discussed above, the incentive effect can dominate the pessimism effect only if persistent restraint puts the person in a situation where he would refrain even in the absence of the incentive effect. In the stationary model, this can happen only if the person is already somewhat addicted, in which case persistent restraint reduces the person's addiction level and thereby reduces the temptation to consume.

The implication of Proposition 2 that sophisticates are more likely than naifs to develop an addiction contradicts the common intuition that harmful addictions are caused by people naively slipping into an unplanned addiction. While we shall in the end vindicate aspects of this intuition, the direct effect of over-optimism is to deter consumption, and therefore the sense in which people naively get addicted is not straightforward. The implication of Proposition 2 that naifs are less likely than sophisticates to quit an established addiction, in contrast, accords well with the common intuition that people "procrastinate" quitting an addiction. Indeed, continuation of an addiction that a person plans to withdraw from is psychologically and mathematically very similar to the type of procrastination discussed in Akerlof (1991) and analyzed in detail in O'Donoghue and Rabin (1999*a*).

It is useful at this point to consider an example:

Example 2: Stationary Linear Model

Suppose $f(k) = -\rho k$ and $g(k) = -(\rho + \sigma)k$.

If $k_1 = 0$, then:
TCs hit always if and only if $\bar{x}_o \geq \frac{\delta\rho}{1-\delta\gamma}$.
Naifs hit always if and only if $\bar{x}_o \geq \min\left\{ \frac{1}{1-\delta+\beta\delta}\frac{\beta\delta\rho}{1-\delta\gamma} \ , \ \frac{\beta\delta(\rho+\sigma)}{1-\delta\gamma} \right\}$.
Sophisticates hit always if and only if $\bar{x}_o \geq \frac{\beta\delta\rho}{1-\delta\gamma}$.

If $k_1 = k^{\max}$, then:
TCs hit always if and only if $\bar{x}_o + \sigma k^{\max} \geq \frac{\delta(\rho+\sigma)}{1-\delta\gamma}$.
Naifs hit always if and only if $\bar{x}_o + \sigma k^{\max} \geq \frac{\beta\delta(\rho+\sigma)}{1-\delta\gamma}$.
Sophisticates hit always if $\bar{x}_o + \sigma k^{\max} \geq \frac{1}{1-\delta+\beta\delta}\frac{\beta\delta(\rho+\sigma)}{1-\delta\gamma}$
and only if $\bar{x}_o + \sigma k^{\max} \geq \frac{\beta\delta(\rho+\sigma)}{1-\delta\gamma}$.

In Example 2, it is easy to see that unaddicted sophisticates are more likely to hit always than unaddicted naifs, but addicted sophisticates are less likely to hit always than addicted naifs. But

more interesting is the fact that simple qualitative comparative statics do not differ across the three types. For all three types, the likelihood of hitting always is increasing in the exogenous temptation $\bar{x}_o$, and decreasing in the patience parameter β, the relevance parameter δ, and the degree of negative internalities ρ. The inherent persistence of addiction γ decreases the likelihood of hitting always when initially unaddicted and increases the likelihood of hitting always when initially addicted, and the degree of habit formation σ increases the likelihood of hitting always when initially addicted.[24] These results illustrate the more general point that simple qualitative comparative-static predictions often cannot distinguish the rational-choice model from our self-control model. Indeed, every comparative-static prediction that we've seen given in support of the rational-addiction model is equally supportive of our self-control model of addiction. We return to this point in Section 6 when we discuss the implications of our model for price comparative statics.

We next turn our attention to the welfare implications of present-biased preferences in the stationary model. There is clearly a popular concern that people are causing themselves severe harm when they develop and maintain harmful addictions. Because rational-choice models of addiction *a priori* assume that people are behaving in their own best interests, they cannot address this concern. Our model, in contrast, shows how present-biased preferences can be a source of over-consumption. Even so, the question remains should we be worried about this over-consumption. Are self-control problems a plausible source of severely harmful addictions, or do they merely lead to minor episodes of suboptimality?

To address this question in a principled way, we first define a notion of "harm". Because we interpret the preference for immediate gratification to be an error, we deem a person's long-run preferences — what the person would prefer if asked at some prior perspective — to be the appropriate preferences for welfare analysis.

Definition 6. A person's **long-run utility** from following strategy α starting from initial addiction level k_1 is $U_1(k_1, \alpha)$. If a person follows a strategy $\alpha \neq \alpha^{tc}$, he suffers **welfare loss** of $WL(k_1, \alpha) = U_1(k_1, \alpha^{tc}) - U_1(k_1, \alpha)$.

[24] For an unaddicted person the degree of habit formation σ plays a role in the propensity to hit only if the person plans to incur withdrawal costs. Because an unaddicted person would incur withdrawal costs only if he planned to hit in the short term and then refrain, in the stationary model σ is relevant only for naifs. The behavior of sophisticates can be annoyingly non-monotonic in some parameters; but recent work by Harris and Laibson (forthcoming) suggests that such non-monotonicities disappear as noise is introduced.

The long-run utility function is the same for all three types. As discussed in Section 2, TC behavior represents how naifs and sophisticates would like to behave if asked from some prior perspective. Hence, the welfare loss for naifs and sophisticates represents their utility loss relative to being able to commit prior to period 1 to some behavior path.[25]

To assess whether present-biased preferences are a plausible source of severe harm, we investigate the maximum welfare loss that a person might suffer as parameters of the model are varied. But since perception-perfect strategies are unchanged by multiplicative transformations of the instantaneous utility function, the magnitude of the welfare loss *per se* is not meaningful. We therefore explore the magnitude of the welfare loss in two ways. First, we express welfare losses in proportion to Δ^H, which is the internality cost from hitting for one period. Calibrationwise, we can then derive the potential harm from plausible self-control problems as a multiple of the internality cost from hitting for one period, which in principle permits one to assess whether the potential harm is empirically "large". Second, we compare the potential harm for sophisticates and naifs to the potential harm for hypothetical *committers* — people with present-biased preferences who can and must commit in period 1 to their desired behavior path. Because committers have the same present-biased preferences as sophisticates and naifs, these results illustrate how the dynamic, one-hit-at-a-time nature of addictive choices contributes to the potential harm suffered by sophisticates and naifs. We let α^{commit} denote the strategy chosen by committers, so welfare losses for committers are $WL(k_1, \alpha^{\text{commit}})$.

Proposition 3 describes the potential welfare losses for committers, sophisticates, and naifs who are initially unaddicted. We restrict attention to the stationary linear model where $f(k) = -\rho k$ and $g(k) = -\phi\rho k$ for some $\phi > 1$, in which case $\Delta^H = \frac{\rho}{1-\delta\gamma}$.[26] We derive the maximum welfare loss when all parameters are fixed except the exogenous temptation $\bar{x}_o$.

[25] An alternative criterion is to measure welfare losses using period-1 preferences instead of long-run preferences, where the benchmark strategy would be α^{commit} rather than α^{tc}. This criterion would yield similar conclusions. A second, and more conservative, alternative is to assume there are no "true preferences", and consider Pareto comparisons (see, e.g., Goldman (1979) and Laibson (1994, 1997)). In our model, sophisticates and naifs follow Pareto-dominated consumption paths whenever they hit always despite preferring in period 1 to never hit, and therefore our welfare approach yields similar conclusions to the more conservative approach. We prefer the long-run-utility criterion because it permits discussion of the *magnitude* of harm.

[26] We believe our welfare conclusions hold qualitatively for more general stationary preferences.

Proposition 3. If $f(k) = -\rho k$ and $g(k) = -\phi\rho k$, then:

$$(1)\ \max_{\bar{x}_o \in \mathbf{R}} \left[WL(0, \alpha^{\text{commit}})\right] = \begin{cases} \delta(1-\beta)\phi\Delta^H & \text{if } 1 < \phi \leq \frac{1}{1-\delta+\beta\delta} \\ \frac{\delta(1-\beta)}{1-\delta+\beta\delta}\Delta^H & \text{if } \phi \geq \frac{1}{1-\delta+\beta\delta}; \end{cases}$$

$$(2)\ \max_{\bar{x}_o \in \mathbf{R}} \left[WL(0, \alpha^s)\right] = \tfrac{\delta(1-\beta)}{1-\delta}\Delta^H \text{ for all } \phi > 1; \text{ and}$$

$$(3)\ \max_{\bar{x}_o \in \mathbf{R}} \left[WL(0, \alpha^n)\right] = \begin{cases} \frac{\delta(1-\phi\beta)}{1-\delta}\Delta^H & \text{if } 1 < \phi \leq \frac{1}{1-\delta+\beta\delta} \\ \frac{\delta(1-\beta)}{1-\delta+\beta\delta}\Delta^H & \text{if } \phi \geq \frac{1}{1-\delta+\beta\delta}. \end{cases}$$

Figure 1 depicts the results from Proposition 3.[27] Part 1 derives the potential harm for hypothetical committers; the results reflect that for low levels of habit formation committers are worst off when they just prefer hitting once to never hitting, whereas for high levels of habit formation committers are worst off when they just prefer hitting always to never hitting. But since committers follow their period-1 desired behavior path, their potential harm is small in the sense that for plausible values of β and δ — reasonably close to 1 — the potential harm is less than Δ^H.[28]

Unlike committers, sophisticates and naifs might hit always despite preferring to never hit, in which case they can suffer significantly larger welfare losses. For sophisticates, such an outcome can arise due to feelings of inevitability: Even when a person prefers never hitting to hitting always, if he believes that he will get addicted in the future he may see no reason to refrain now. Part 2 of Proposition 3 shows that such reasoning can lead to severe harm for sophisticates in the sense that for plausible values of β and δ their potential harm can be much larger than Δ^H, and is a multiple — greater than or equal to $\frac{1-\delta+\beta\delta}{1-\delta}$ — of the potential harm for committers.[29] The potential harm for sophisticates is independent of the degree of habit formation. If harmful addictions are caused by sophistication and feelings of inevitability, the degree of habit formation is irrelevant to the question of how much damage a person might do to himself when he chooses to develop an addiction.

Naifs might hit always despite preferring to never hit when they repeatedly plan on short-term consumption and end up with long-term consumption. But the habit-forming property of addictive

[27] Figure 1 is drawn to scale for the case $\beta = .8$ and $\delta = .95$.

[28] It is instructive to calibrate our welfare results for some specific values of β and δ. We shall (somewhat arbitrarily) focus on two cases: (I) $\beta = .8$ and $\delta = .95$, and (II) $\beta = .99$ and $\delta^{365} = .95$. Case I is meant to be plausible when the length of a period is on the order of one year, and Case II is meant to be plausible when the length of a period is on the order of one day. The potential harm for committers is at most $.23\Delta^H$ in Case I, and $.01\Delta^H$ in Case II.

[29] Calibrationwise, and using the cases from footnote 28, the potential harm for sophisticates is $3.80\Delta^H$ in Case I and $71.15\Delta^H$ in Case II. These values are at least 16.5 times and 7115 times the potential harm for committers, respectively.

goods tends to deter short-term consumption because short-term consumption creates unwanted future withdrawal costs. Indeed, in a stationary model, if the product is sufficiently habit-forming, short-term consumption is never desirable. Part 3 of Proposition 3 reflects this intuition by establishing that if the product is sufficiently habit-forming — ϕ is large enough — the potential harm for naifs is identical to that for committers. For smaller degrees of habit formation, the potential harm for naifs is a multiple of that for committers, although it is smaller than that for sophisticates (see Figure 1).[30]

Our welfare results in Proposition 3 correspond to our earlier behavioral conclusion that sophisticates are more likely than naifs to develop an addiction. But naifs are more likely than sophisticates to maintain an established addiction. To investigate the potential harm that naifs could suffer from maintaining an established addiction, Proposition 4 characterizes how the potential harm for hypothetical committers and for naifs depends on the initial addiction level.[31]

Proposition 4. Let $f(k) = -\rho k$, $g(k) = -\phi\rho k$, and fix $\phi > \frac{1}{1-\delta+\beta\delta}$. There exists $k^* \in (0, (1-\delta+\beta\delta)k^{\max})$ such that:

$$(1)\ \max_{\bar{x}_o \in \mathbf{R}} \left[WL(k_1, \alpha^{\text{commit}})\right] = \begin{cases} \frac{\delta(1-\beta)}{1-\delta+\beta\delta}\Delta^H \left[1 + \frac{k_1}{k^{\max}}(\phi - 1)\right] & \text{if } k_1 \leq k^* \\ \delta(1-\beta)\phi\Delta^H & \text{if } k_1 \geq k^*; \end{cases}$$

$$(2)\ \max_{\bar{x}_o \in \mathbf{R}} \left[WL(k_1, \alpha^n)\right] = \begin{cases} \frac{\delta(1-\beta)}{1-\delta+\beta\delta}\Delta^H \left[1 + \frac{k_1}{k^{\max}}(\phi - 1)\right] & \text{if } k_1 \leq k^* \\ \frac{\delta(1-\beta)\phi}{1-\delta}\Delta^H - \left(1 - \frac{k_1}{k^{\max}}\right)\frac{\delta(\phi-1)}{1-\delta}\Delta^H & \text{if } k_1 \geq k^*. \end{cases}$$

Proposition 4 fixes the degree of habit formation to be sufficiently large that an unaddicted naif would not suffer severe harm. If a person is sufficiently unaddicted, then short-term consumption is not desirable for the reasons outlined above, and the potential harm for naifs is identical to that for committers. But as the person becomes more addicted, a new force becomes important. Unlike an unaddicted person, an addicted person who plans to eventually quit must incur withdrawal costs from past consumption. While current consumption still creates additional unwanted future withdrawal costs, which tends to deter current consumption, current consumption also delays the

[30] Calibrationwise, and using the cases from footnote 28, the potential harm for naifs is identical to that for committers if $\phi \geq 1.23$ in Case I and if $\phi \geq 1.01$ in Case II.

[31] When $k_1 > 0$, solving for potential welfare losses for sophisticates is a mess, and some preliminary calculations suggested that there are no new insights.

withdrawal costs from past consumption, which tends to encourage current consumption. For a sufficiently addicted person, the latter effect dominates, and therefore the potential harm for naifs can be a multiple of the potential harm for committers. Indeed, for $k_1 = k^{\max}$, the potential harm for naifs is $\frac{1}{1-\delta}$ times the potential harm for committers.[32]

Hence, in our stationary model, there are two sources of severe harm. To the extent that people are sophisticated, they may suffer severe harm when they develop an addiction due to feelings of inevitability. To the extent that people are naive, they may suffer severe harm when they procrastinate quitting an established addiction. While this latter source is relatively unimportant in the stationary model — because naifs would never develop the addiction in the first place — it becomes crucial in the more realistic non-stationary case we consider next.

5. Youthful Preferences

In Section 4 we make the unrealistic assumption that the instantaneous utility function is constant over time. It is more likely that the temptation to consume may vary over time in systematic or random ways. In this section, we explore the implications of one particular type of non-stationarity wherein the temptation to hit is larger earlier in life. Formally:

Youthful Preferences:

$$u_t(a,k) \equiv \begin{cases} x_t + f(k) & \text{if } a = 1 \\ y_t + g(k) & \text{if } a = 0 \end{cases}$$

where $\bar{x}_1 \geq \bar{x}_2 \geq ... \geq \bar{x}_M = \bar{x}_{M+1} = ... = \bar{x}_T$.

Because the temptation to hit in period t is $h_t(k) \equiv [x_t - y_t] + [f(k) - g(k)]$, this assumption implies that while the endogenous temptation $f(k) - g(k)$ is independent of the person's age, the exogenous temptation $\bar{x}_t \equiv x_t - y_t$ decreases as the person ages. Our formal results also assume that a person eventually "matures" in that beginning in some period $M < \infty$ preferences become stationary; this is a vacuous limitation for $T < \infty$, but is a restriction in the infinite-horizon case on which we focus.[33]

[32] Also note that the larger the degree of habit formation, the larger is the potential harm for naifs. Intuitively, the more habit-forming is the product, the more prone are naifs to procrastinate quitting.

[33] The assumption of a maturity date is used only for results concerning sophisticates; we conjecture that our results and intuitions for sophisticates hold more generally. As in the stationary model, for sophisticates we restrict attention to perception-perfect strategies for the infinite horizon that correspond to the unique perception-perfect strategy for some long, finite horizon, where we fix a finite maturity date M and let T become large.

Youthful instantaneous utilities reflect forces such as peer pressure and intrinsic biological factors that lead most people to face larger temptations while young than they do later in life. Youthful instantaneous utilities might also reflect the effects of a traumatic life event, such as a divorce, loss of a job, or death of a loved one: After a traumatic event, a person may have an increased desire to consume an addictive product for some duration, but eventually the desire to consume returns to more normal levels. Most importantly, youthful instantaneous utilities permit the plausible assumption that an addictive product is intrinsically appealing early in life but not later in life.

Our first goal in this section is to explore how our conclusions from the stationary model change in the more realistic youthful model. We begin with some preliminary results concerning how the three types behave in the youthful model.

Lemma 6. Under youthful instantaneous utilities and $T = \infty$:

(1) $\bar{k}_t^{tc} \leq \bar{k}_{t+1}^{tc}$ for all t,

(2) $\bar{k}_t^{n} \leq \bar{k}_{t+1}^{n}$ for all t, and

(3) If $\bar{x}_M \geq \beta\delta\Delta^H$, then $\alpha^s(k,t) = 1$ for all k and t. If $\bar{x}_M < \beta\delta\Delta^H$, then there exists $k' > 0$ such that for all $t \geq M$, $\alpha^s(k,t) = 0$ for all $k < k'$.

Parts 1 and 2 of Lemma 6 establish that for both TCs and naifs, the cutoff addiction level above which the person hits is smaller in earlier periods. The intuition is simple: In the youthful model the temptation to hit is larger in earlier periods, and since TCs and naifs plan to follow optimal consumption paths, they are each more likely to hit in earlier periods. An immediate implication of this result is that both TCs and naifs hit first and refrain later: Starting from any situation, they either never hit again, hit for a finite number of periods and then never hit again, or hit always.

Part 3 of Lemma 6 establishes that, in contrast to the stationary model where the crucial question for sophisticates is whether they feel that addiction is inevitable *in period 1*, the crucial question in the youthful model is whether they feel that addiction is inevitable *at maturity*, which holds if $\bar{x}_M \geq \beta\delta\Delta^H$. If a sophisticate views addiction as inevitable at maturity, then he clearly hits throughout his youth when the temptation to hit is even larger. If, in contrast, the sophisticate would refrain once mature if sufficiently unaddicted, then he may refrain in his youth as well.

Proposition 5 shows that, as in the stationary model, the inevitability condition determines when the incentive effect is operative, and hence determines when sophisticates are less or more prone

to hit than naifs.

Proposition 5. Under youthful instantaneous utilities and $T = \infty$:

(1) If $\bar{x}_M \geq \beta\delta\Delta^H$, then $\alpha^s(k,t) \geq \alpha^n(k,t)$ for all k and t, and

(2) If $\bar{x}_M < \beta\delta\Delta^H$, then $\alpha^s(k,t) \leq \alpha^n(k,t)$ for all k and t.

Proposition 5 implies that under the plausible assumption that an addictive product eventually loses its intrinsic appeal, sophisticates are never more prone to hit than naifs. Moreover, unlike in the stationary model, sophisticates can be strictly less prone to hit than naifs even when unaddicted, as illustrated in Example 3.

Example 3: Suppose $\delta = .9$, $\beta = .5$, and $\gamma = 0$. Let $f(k) = -3k$, $g(k) = -12k$, and suppose $\bar{x}_1 = 8$ and $\bar{x}_t = 1$ for all $t \geq 2$. Then starting from $k_1 = 0$:

(1) TCs never hit,

(2) Naifs hit always, and

(3) Sophisticates never hit.

In Example 3, once they reach maturity in period 2, both naifs and sophisticates will refrain forever after if and only if they are unhooked at $t = 2$. Hence, the incentive effect becomes important for preventing unwanted addictions. Sophisticates recognize that indulging in the youthful temptation would lead to a lifetime of hitting, and so refrain to induce good behavior in their maturity. Naifs, in contrast, think in their youth that they can indulge in the large youthful temptation and later quit, but this unfortunately leads to a lifetime of hitting. Example 3 illustrates an important difference between the stationary and youthful models: In the youthful model, persistent restraint can reduce the temptation to hit even for an unaddicted person, and hence the incentive effect can dominate the pessimism effect even for an unaddicted person.

Hence, stationary models may make overly pessimistic predictions for sophisticates, and overly optimistic predictions for naifs. To allow a more systematic analysis of these points, we define a formal sense in which a given youthful environment is comparable to a stationary environment. We define a *youthful rotation* to be a transformation of stationary environment into a youthful environment that holds constant both the period-1 utility from hitting always and the period-1 utility from never hitting.

Definition 7. Consider stationary instantaneous utility function

$$u_t(a,k) \equiv \begin{cases} x_o + f(k) & \text{if } a = 1 \\ g(k) & \text{if } a = 0 \end{cases}$$

and youthful instantaneous utility function

$$\hat{u}_t(a,k) \equiv \begin{cases} x_t + f(k) & \text{if } a = 1 \\ g(k) & \text{if } a = 0 \end{cases}$$

for some $x_1 \geq x_2 \geq ... \geq x_M = x_{M+1} = ... = x_T$. We say that $\hat{u}_t$ is a **youthful rotation** of u_t if $x_1 > x_o$ and $x_1 + \beta \sum_{t=2}^{T} \delta^{t-1} x_t = x_o + \beta \sum_{t=2}^{T} \delta^{t-1} x_o$.[34]

Because a youthful rotation makes early-life hitting more attractive and late-life hitting less attractive, a youthful rotation can clearly cause a person to switch from never hitting or always hitting to hitting only in his youth. The more interesting question is whether a youthful rotation can cause a person to switch from hitting always to never hitting and vice-versa.

Proposition 6. Suppose $\hat{u}_t$ is a youthful rotation of u_t, and let $\mathbf{a}^{tc}$, $\mathbf{a}^n$, $\mathbf{a}^s$, $\widehat{\mathbf{a}}^{tc}$, $\widehat{\mathbf{a}}^n$, and $\widehat{\mathbf{a}}^s$denote the perception-perfect behavior paths given $k_1 = 0$ under u_t and $\hat{u}_t$. Then:

(1) $\mathbf{a}^{tc} = (1,1,...)$ implies $\widehat{\mathbf{a}}^{tc} \neq (0,0,...)$, and $\mathbf{a}^{tc} = (0,0,...)$ implies $\widehat{\mathbf{a}}^{tc} \neq (1,1,...)$;

(2) $\mathbf{a}^n = (1,1,...)$ implies $\widehat{\mathbf{a}}^n \neq (0,0,...)$; and

(3) if $\mathbf{a}^s = (0,0,...)$ and $\widehat{\mathbf{a}}^s = (1,1...)$, then $\mathbf{a}^n = (0,0,...)$ and $\widehat{\mathbf{a}}^n = (1,1,...)$.

Part 1 establishes that for TCs a youthful rotation can neither cause a switch from hitting always to never hitting nor vice-versa. Because a youthful rotation does not change the utility from these two options, these results follow from a simple application of revealed preference. Part 2 establishes that for naifs a youthful rotation cannot cause a switch from hitting always to never hitting for essentially the same revealed-preference reason. But a youthful rotation *can* cause naifs to switch from never hitting to hitting always, because for a stationary model in which naifs never hit, a youthful rotation can make them plan to hit in their youth and later quit, and once they have become somewhat hooked on the product they might never quit.

For sophisticates, a youthful rotation can cause a switch in either direction. Youthful rotations can cause sophisticates to switch from never hitting to hitting always by creating an irresistible temptation to hit during youth, after which it may be worthwhile to continue hitting. We don't

[34] Since TCs have $\beta = 1$, for TCs $\hat{u}_t$ is a youthful rotation of u_t if $x_1 > x_o$ and $\sum_{t=1}^{T} \delta^{t-1} x_t = \sum_{t=1}^{T} \delta^{t-1} x_o$.

believe this is a particularly important intuition, and moreover Part 3 of Proposition 6 establishes that sophisticates switch from never hitting to hitting always only in situations where naifs also switch from never hitting to hitting always. We believe the more important intuition is that a youthful rotation can cause sophisticates to switch from hitting always to never hitting. In a stationary model, sophisticates sometimes hit always even though they would prefer to never hit because they believe late-life hitting is inevitable. By decreasing the temptation later in life, a youthful rotation may eliminate the inevitability of addiction, and therefore enable sophisticates to refrain always.

Proposition 6 implies that youthful rotations can have opposite implications for naifs and sophisticates. Example 4 illustrates the opposite implications can arise *for the same youthful rotation*.

Example 4: Suppose $\delta = .9$, $\beta = .5$, and $\gamma = 0$. Let $f(k) = -3k$, $g(k) = -12k$, and suppose $k_1 = 0$:

(1) If $\bar{x}_t = 1.7$ for all $t \geq 1$, then sophisticates hit always whereas naifs never hit; and

(2) If $\bar{x}_1 = 8$ and $\bar{x}_t = 1$ for all $t \geq 2$, then naifs hit always whereas sophisticates never hit.

We next explore the welfare implications of present-biased preferences in the youthful model. Again, we are interested in whether self-control problems represent a plausible source of severely harmful addictions, and hence focus on maximum possible welfare losses. Proposition 7 characterizes the potential harm for hypothetical committers and naifs who are initially unaddicted:

Proposition 7. If $f(k) = -\rho k$ and $g(k) = -\phi\rho k$, then:

(1) $\max_{(\bar{x}_1,\bar{x}_2,\ldots)\in\mathbf{R}^\infty} \left[WL(0,\alpha^{\text{commit}})\right] = \delta(1-\beta)\phi\Delta^H$ for any $\phi > 1$; and

(2) $\max_{(\bar{x}_1,\bar{x}_2,\ldots)\in\mathbf{R}^\infty} \left[WL(0,\alpha^{n})\right] = \frac{\delta(1-\beta)\phi\Delta^H}{1-\delta}$ for any $\phi > 1$.

The more realistic youthful environment is problematic for naifs because they can be tempted in their youth to acquire an addiction that they delay quitting for the rest of their lives. Indeed, Proposition 7 reveals two senses in which naive self-control problems may be a plausible source of severely harmful addictions in this environment. First, a comparison of naifs to committers in the youthful model reveals that the potential harm for naifs is $\frac{1}{1-\delta}$ times the potential harm for committers for any degree of habit formation. Second, a comparison of naifs in the stationary vs. youthful model — comparing Propositions 3 and 7 — reveals that the potential harm for naifs is much higher in the youthful model (see Figure 2).

Unfortunately, we have not found any general welfare results for sophisticates. But we can describe the ways in which sophisticates might harm themselves and the likely implications. In situations where there is an inevitability to addiction at maturity, then sophisticates can suffer welfare losses in much the same way as they do in the stationary model — because they develop a lifelong addiction due to a lifelong feeling of inevitability. Clearly the potential harm from such an addiction can be just as large as — but no larger than — that for the stationary model.

If, in contrast, there is no inevitability at maturity, sophisticates can suffer welfare losses of a different form. First, sophisticates might suffer welfare losses because they hit too much during their youth, as illustrated in Example 5.

Example 5: Suppose $\delta = .9$, $\beta = .6$, and $\gamma = 0$. Let $f(k) = -20k$, $g(k) = -25k$, and suppose $M > 2$, $\bar{x}_t = 15$ for all $t < M$, and $\bar{x}_t = 5$ for all $t \geq M$. Then starting from $k_1 = 0$:

(1) TCs never hit,

(2) Sophisticates hit for the first $M - 1$ periods and then refrain thereafter.

In Example 5, sophisticates correctly predict that they will refrain once mature no matter what they do during their youth. As a result, some indulgence in their youth is "safe" in the sense that it won't cause a lifelong addiction, and in this example sophisticates end up indulging throughout their youth. Such over-indulgence during one's youth can cause welfare losses because sophisticates give too little weight to the eventual withdrawal costs.

The second way in which sophisticates can hurt themselves that cannot arise in the stationary model is that they might *under*-consume in their youth as a means of preempting over-consumption at maturity. Example 6 illustrates this possibility:

Example 6: Suppose $\delta = .9$, $\beta = .9$, and $\gamma = .999$. Let $f(k) = -k$, $g(k) = -2.5k$, and suppose $\bar{x}_1 = 24$, $\bar{x}_2 = \bar{x}_3 = 18.6$, and $\bar{x}_t = 0$ for all $t \geq 4$. Then starting from $k_1 = 0$:

(1) TCs hit in period 1 and then refrain thereafter,

(2) To preempt consumption in periods 2 and 3, sophisticates never hit.

In Example 6, both TCs and sophisticates would like to hit in period 1 when the enjoyment from hitting is very high, and then never hit again. TCs follow precisely this plan. But sophisticates recognize that hitting in period 1 would lead to unwanted further consumption in periods 2 and 3, and therefore refrain in period 1. Examples 5 and 6 illustrate that even when the incentive effect

is operative, sophisticates can still suffer welfare losses. But our impression from the examples we have worked through is that under the plausible assumption that addiction is not inevitable at maturity, sophisticates are much less prone to suffer severe welfare losses in the youthful model than in the stationary model.

The youthful model can also be used to shed light on an issue that we feel is misleadingly discussed in the rational-addiction literature. Using a stationary model, Becker and Murphy (1988) describe how it can be optimal for a person to *maintain* an established harmful addiction, but their use of steady-state analysis prevents them from analyzing why a person would choose to *develop* the harmful addiction in the first place. They suggest that events such as youth, divorce, and death of a loved one are plausible sources of harmful addictions. Our youthful model allows us to directly investigate this hypothesis, because we can ask whether and by how much traumatic events can harm a person by leading him to develop an addiction.

Suppose that, absent a traumatic event, a person has a stationary, linear instantaneous utility function with no intrinsic desire to hit — that is, for all t

$$u_t(a,k) \equiv \begin{cases} -\rho k & \text{if } a = 1 \\ -(\rho+\sigma)k & \text{if } a = 0. \end{cases}$$

With such instantaneous utilities, an unaddicted person would never hit —regardless of his type — but a person with an established addiction might. Suppose that a traumatic event increases the temptation to consume for N periods, and in particular makes refraining more painful. Formally, we assume that the person faces instantaneous utility function

$$u_t(a,k) \equiv \begin{cases} -\rho k & \text{if } a = 1 \\ -y_t - (\rho+\sigma)k & \text{if } a = 0 \end{cases}$$

where $y_1 \geq y_2 \geq ... \geq y_N > 0$ and $y_t = 0$ for all $t > N$.

For all three types, a traumatic event could of course cause a lifelong addiction. This qualitative aspect of Becker and Murphy's story is obviously correct. But how harmful could such an addiction be? For TCs and sophisticates, there is a sense in which the answer is not very harmful:

Proposition 8. Consider an N-period traumatic event on an otherwise untempting product.

(1) $\min_{(\rho,\sigma)\in\mathbf{R}^2_+} \left[U_1(0,\alpha^{TC})\right] \quad = \quad -\left(\sum_{t=1}^{N} \delta^{t-1} y_t\right)$; and

(2) $\min_{(\rho,\sigma)\in\mathbf{R}^2_+} \left[U_1(0,\alpha^{s})\right] \quad \geq \quad -\left(\frac{1}{\beta}\sum_{t=1}^{N} \delta^{t-1} y_t\right)$.

Absent the traumatic event, all three types would never hit and therefore experience long-run utility $U_1(0, \alpha^i) = 0$. Proposition 8 therefore describes by how much a person might be hurt by the traumatic event. Part 1 establishes that the most a TC might be hurt is by the present discounted sum of the pain from not consuming during the traumatic event. Intuitively, if a traumatic event causes a TC to develop an addiction, then at the moment the traumatic event occurs developing the addiction is better than never hitting.[35] Part 2 establishes that a similar result holds for sophisticates. Given no intrinsic desire to hit absent the traumatic event, a sophisticate does not view a lifelong addiction as inevitable — if he can reach the end of the traumatic event sufficiently unhooked, then he will refrain thereafter. This knowledge limits how much the sophisticate can be hurt by the traumatic event, because if a lifelong addiction is too harmful then he'll make sure to reach the end of the traumatic event sufficiently unhooked.[36]

By contrast, a traumatic event can lead naifs to develop a lifelong addiction well out of proportion from the pain of the traumatic event itself. A naif may think it is safe to hit during the traumatic event because he'll quit once sobriety becomes less painful. But if consuming during the traumatic event gets him sufficiently addicted, the naif procrastinates quitting and as a result suffers large harm. The cleanest case to illustrate this point is for a one-period traumatic event when $\gamma = 0$.

Example 7: Consider a one-period traumatic event and suppose $\gamma = 0$.

(1) $\min_{(\rho,\sigma)\in\mathbf{R}^2_+} \left[U_1(0, \alpha^{TC})\right] = -(y_1)$;

(2) $\min_{(\rho,\sigma)\in\mathbf{R}^2_+} \left[U_1(0, \alpha^{s})\right] = -\left(\frac{1}{\beta}y_1\right)$; and

(3) $\min_{(\rho,\sigma)\in\mathbf{R}^2_+} \left[U_1(0, \alpha^{n})\right] = -\left(\frac{1}{\beta}y_1 + \frac{\delta}{1-\delta}\frac{1-\beta}{\beta}y_1\right)$.

The source of harm for naifs in Example 7 is not the traumatic event *per se*, but rather that naifs fail to quit the addiction caused by the traumatic event. Indeed, the additional harm that naifs might suffer relative to sophisticates is essentially the maximum welfare loss that they might suffer from not quitting an established addiction. As long as the future holds enough relevance — δ is close enough to 1 — this latter source of harm can be many times the pain of the traumatic event itself.[37]

[35] For TCs, a second qualitative feature of "traumatic-event-caused" addictions is that the person consciously chooses to develop the addiction at the moment the traumatic event occurs. While traumatic events may lead some people to consciously choose a lifelong addiction — as in the movie *Leaving Las Vegas* — we suspect that many such addictions are not intentional.

[36] Proposition 8 provides a lower bound on sophisticates' utility, but for many $(y_1, ..., y_N)$ this bound cannot be achieved. Hence, Proposition 8 *over*states by how much sophisticates can be hurt.

[37] More generally, a traumatic event can cause a severely harmful addiction for naifs as long as it gets them sufficiently addicted that they procrastinate quitting.

Hence, our model suggests that traumatic events may be a plausible source of severely harmful addictions for naifs, but not for TCs and sophisticates.

6. Price Effects

In this section, we examine the effects of price on consumption. Because our analysis of prices is conducted within the confines of our binary-choice model, it is crude in a number of ways. But we feel it captures some important intuitions that would hold in a more general model. Our main goal is to provide some intuition for why existing empirical evidence often invoked as support for the rational-choice (exponential) model of addiction may in fact be more supportive of a self-control model of addiction.

To introduce prices into our model of instantaneous utilities, we suppose that in period t the person consumes the addictive product and "other goods". We assume that the person's income in period t is Y_t, and that he cannot borrow or save. We assume the price of other goods is normalized to one, and that the price of the addictive product in period t is p_t. Hence, in period t, if the person refrains then he consumes quantity Y_t of other goods, and if he hits then he consumes quantity $Y_t - p_t$ of other goods. Assuming that utility from the addictive product is stationary and that utility from other goods is stationary, linear, and additively separable from utility from the addictive product, we can re-write the person's instantaneous utility function as:

$$u_t(a,k) \equiv \begin{cases} f(k) + [Y_t - p_t] & \text{if } a = 1 \\ g(k) + [Y_t] & \text{if } a = 0. \end{cases}$$

Because of the discreteness of our model, there is limited scope for studying marginal price changes. We can, however, analyze marginal changes in the cutoff addiction level for which a person consumes. That is, Lemma 1 implies that for any price vector $(p_1, ..., p_T)$, both TCs and naifs follow a cutoff strategy, and we analyze how price changes affect the period-1 cutoffs $\bar{k}_1^{tc}$ and $\bar{k}_1^{n}$. Both for simplicity (because sophisticates need not follow a cutoff rule) and because we believe naivete is the more empirically relevant case, we confine our analysis to TCs and naifs.

We suppose that there is initially a fixed price $\bar{p}$ — i.e., $p_t = \bar{p}$ for all t — and consider three price comparative statics: an immediate permanent price change — a change in $\bar{p}$ — which we denote by $d\bar{k}_1^i/d\bar{p}$; an immediate temporary price change — a change in p_1 holding $p_t = \bar{p}$ for all $t \neq 1$ — which we denote by $d\bar{k}_1^i/dp_1$; and an expected future temporary price change — a change

in a future price p_τ holding $p_t = \bar{p}$ for all $t \neq \tau$ — which we denote by $d\bar{k}_1^i/dp_\tau$.[38]

It is straightforward to derive that all qualitative price comparative statics are the same: For both TCs and naifs, a price increase — whether it be permanent, immediate temporary, or future temporary — causes a person's cutoff $\bar{k}_1$ to increase, which means the person is less prone to consume in period 1.[39] Much as we discussed for Example 2, simple comparative static results are the same for TCs and naifs. Indeed, Gruber and Koszegi (2000) investigate price comparative statics in a continuous-choice model (with additional assumptions about functional forms), and reach the same conclusion. Hence, the most common test of the rational-choice model of addiction — whether current consumption depends on future prices — does not test whether people have self-control problems. Our point in this section, however, is that if one looks more carefully at these empirical results, calibrationwise they may be more supportive of a self-control model of addiction.

Because in our model absolute price comparative statics are not meaningful, we focus on relative price comparative statics — e.g., the impact of a permanent price change relative to the impact of a temporary price change. Proposition 9 derives some relative price comparative statics. The values $k^*(\beta, \bar{p})$ and $\tilde{k}(\beta, \bar{p})$ are the analogues of $k^*(\beta)$ and $\tilde{k}(\beta)$ in Section 4 (for a fixed price $\bar{p}$ the price model is equivalent to the stationary model).

Proposition 9. Suppose $p_t = \bar{p}$ for all t, and that $\bar{k}_1^{tc}, \bar{k}_1^n \in (0, k^{\max})$. Then:

(1) For people with time-consistent preferences,

$$\frac{d\bar{k}_1^{tc}}{dp_\tau} \Big/ \frac{d\bar{k}_1^{tc}}{dp_1} = \delta^{\tau-1} \qquad \text{and} \qquad \frac{d\bar{k}_1^{tc}}{d\bar{p}} \Big/ \frac{d\bar{k}_1^{tc}}{dp_1} = \frac{1}{1-\delta}.$$

(2) For people with present-biased preferences who are naive:

(a) If the initial price $\bar{p}$ is such that $k^*(\beta, \bar{p}) < \tilde{k}(\beta, \bar{p})$,

$$\frac{d\bar{k}_1^{n}}{dp_\tau} \Big/ \frac{d\bar{k}_1^{n}}{dp_1} = \beta\delta^{\tau-1} \qquad \text{and} \qquad \frac{d\bar{k}_1^{n}}{d\bar{p}} \Big/ \frac{d\bar{k}_1^{n}}{dp_1} = 1 + \frac{\beta\delta}{1-\delta}.$$

(b) If the initial price $\bar{p}$ is such that $k^*(\beta, \bar{p}) > \tilde{k}(\beta, \bar{p})$,

$$\frac{d\bar{k}_1^{n}}{dp_\tau} \Big/ \frac{d\bar{k}_1^{n}}{dp_1} = 0 \qquad \text{and} \qquad \frac{d\bar{k}_1^{n}}{d\bar{p}} \Big/ \frac{d\bar{k}_1^{n}}{dp_1} = 1.$$

[38] This technique is essentially the same as that used in Becker and Murphy (1988) and Gruber and Koszegi (2000).

[39] There is one caveat: In situations where they don't expect to consume in the future, naifs have no reaction to future price changes. But this no-reaction result is an artifact of our discrete-choice model. All relevant price comparative statics are derived in the proof of Proposition 9 below.

Part 1 presents comparative statics for TCs. For a fixed price $\bar{p}$, TCs hit in period 1 if and only if a lifetime of hitting is preferred to a lifetime of restraint. Hence, a price change affects behavior only to the extent that it makes a lifetime of hitting look more or less worthwhile, and therefore the relative price comparative statics are equal to the relative amounts by which current, future, and permanent price changes affect the cost of hitting always.

Part 2 presents comparative statics for naifs. Lemma 5 establishes that for a fixed price $\bar{p}$, the cutoff for naifs can be either the addiction level at which the person is indifferent between never hitting and hitting always or the addiction level at which the person is indifferent between never hitting and hitting once. In the former case, where naifs like TCs hit in period 1 if and only if a lifetime of hitting is preferred to a lifetime of restraint, part 2a establishes that the relative comparative statics are similar to those for TCs, differing only to the extent that naifs discount future periods by the factor β. In the latter case, naifs hit in period 1 if and only if hitting once is preferred to never hitting, and hence a price change affects behavior only to the extent that it affects the utility of hitting once. Part 2b establishes that this case yields very different comparative statics. In particular, because an individual with k_1 near $\bar{k}_1^n$ does not plan to consume in the future, future prices do not affect the cutoff (at least for small price changes).

Table 1 explores calibrationwise — for some reasonable parameter values of β and δ — what these relative price comparative statics should be.

Table 1: (Unanticipated) Price Comparative Statics in Our Model

	Elasticity	TCs	Naifs if $k^*(\beta,\bar{p}) < \tilde{k}(\beta,\bar{p})$	Naifs if $k^*(\beta,\bar{p}) > \tilde{k}(\beta,\bar{p})$
$\beta = .9 \quad \delta = .95$	$\frac{d\bar{k}_1^i}{dp_2} \Big/ \frac{d\bar{k}_1^i}{dp_1}$	0.95	0.86	0
	$\frac{d\bar{k}_1^i}{d\bar{p}} \Big/ \frac{d\bar{k}_1^i}{dp_1}$	20.00	18.10	1
$\beta = .8 \quad \delta = .95$	$\frac{d\bar{k}_1^i}{dp_2} \Big/ \frac{d\bar{k}_1^i}{dp_1}$	0.95	0.76	0
	$\frac{d\bar{k}_1^i}{d\bar{p}} \Big/ \frac{d\bar{k}_1^i}{dp_1}$	20.00	16.20	1
$\beta = .9 \quad \delta = .9$	$\frac{d\bar{k}_1^i}{dp_2} \Big/ \frac{d\bar{k}_1^i}{dp_1}$	0.90	0.81	0
	$\frac{d\bar{k}_1^i}{d\bar{p}} \Big/ \frac{d\bar{k}_1^i}{dp_1}$	10.00	9.05	1

Table 1 reveals that for TCs with a plausible yearly discount factor δ, temporary price changes now vs. next period have similar effects on the period-1 cutoff, whereas a permanent price change has a much larger effect. Similar conclusions hold for naifs when their cutoff is the addiction level at which they are indifferent between never hitting and hitting always. But when the cutoff for naifs is the addiction level at which they are indifferent between never hitting and hitting once, a very different pattern emerges. Because future price changes do not affect the cutoff, a temporary price change next period has a very small effect relative to a temporary price change now, and a permanent price change and an immediate temporary price change have identical effects.

While these comparative statics are artificially extreme due to our crude model, they reflect the more general point that, because naifs underestimate how much they'll consume in the future, future prices matter much less for naifs than for TCs. To further illustrate this point, consider a person who has a pack-a-day smoking habit. If this person is time-consistent, he smokes one pack a day because doing so is optimal, and moreover he plans to be smoking one pack a day. Because a temporary change in the price of cigarettes has only a small effect on the lifetime cost of his chosen behavior, whereas a permanent change in the price of cigarettes significantly changes the lifetime cost of smoking one pack a day, for TCs permanent price changes should have significantly larger effects than temporary price changes. Suppose instead that the person has present-biased preferences and is naive, in which case he may be smoking one pack a day his entire life *not* because he finds a lifetime pack-a-day habit optimal, but rather because he always plans to smoke one pack a day for a short while and then quit. For such a person, the only relevant prices are those for the near future. Hence, a permanent price change and an immediate temporary price change should have similar effects on consumption.

In this light, we now reinterpret the empirical literature on rational addiction. The main empirical finding is that consumption of addictive products depends on past and future prices, suggesting both that the products are indeed addictive and that people are forward-looking and take into account how current consumption affects future well-being.[40] But a puzzle in this literature is that temporary price changes and permanent price changes have similar effects on consumption. As an example, Table 2 presents the price elasticities derived in Becker, Grossman, and Murphy's 1994 study of cigarette consumption, where $\varepsilon(x, y)$ is the point elasticity of variable x with respect to variable y.

[40] We emphasize again that while these results are consistent with the rational-choice model of addiction, they are also consistent with our model of addiction.

Table 2: (Unanticipated) Price Elasticities from
Becker, Grossman, and Murphy (1994, Table 4)

Elasticity	Model (i)	Model (ii)	Model (iii)	Model (iv)
$\varepsilon(C_t, p_t)$	-0.349	-0.322	-0.316	-0.262
$\varepsilon(C_t, p_{t+1})$	-0.050	-0.084	-0.058	-0.068
$\varepsilon(C_t, \bar{p})$	-0.407	-0.436	-0.387	-0.355
$\varepsilon(C_t, p_{t+1})/\varepsilon(C_t, p_t)$	0.14	0.26	0.18	0.26
$\varepsilon(C_t, \bar{p})/\varepsilon(C_t, p_t)$	1.17	1.35	1.22	1.35

Becker, Grossman, and Murphy recognize the puzzle, noting that the regression results reported above imply absurd yearly discount rates ranging from 56.3% to 222.6%. They conclude that the data is too coarse to identify the discount rate. But Table 2 reveals that the relative comparative statics look very much like those for naifs in our model — that is, close to zero and one rather than one and much larger than one. Hence, our self-control model of addiction suggests an alternative explanation for the puzzling empirical results: While the regression results imply absurd discount rates *under the maintained hypothesis of time consistency*, they may be quite consistent with plausible discount rates once one permits that people might have a small self-control problem about which they are naive. Hence, calibrationwise the existing empirical literature on rational addiction may be more supportive of a self-control model of addiction than of the fully rational model of addiction.[41]

7. Discussion

We conclude by discussing why our self-control model of addiction is an improvement relative to the rational-choice model of addiction, and some general lessons to be gleaned from our analysis.

The most obvious advantage of our model is simple realism. While economists have become habituated to the exponential-discounting model, the evidence overwhelmingly supports the hy-

[41] Of course, we don't want to sell this point too strongly. One of our goals for the future is to develop a self-control model that can be more readily taken to the data so as to further test this hypothesis.

pothesis that people have present-biased preferences.[42] Of course, because in many domains self-control problems likely have marginal effects, time consistency is often a useful approximation to the more realistic model of present-biased preferences. But addiction is a realm where intuition suggests self-control problems matter a lot, and hence the obviously appropriate null hypothesis for studies of addiction should be that self-control problems matter.

Related to the issue of realism, we predict that our self-control model of addiction — especially when it incorporates an element of naivete — will be better calibrated than the rational-choice model, and hence make sounder *quantitative* predictions.[43] We have already seen an example of this point in Section 6: Under the maintained hypothesis of time consistency, the empirical addiction results imply absurd discount rates, whereas the same results are consistent with plausible discount rates and a small present bias about which the person is naive. More generally, we suspect that if one were to estimate discount rates and the various properties of an addictive product — e.g., addictiveness, degree of negative internalities, etc. — the behavior of addicts just wouldn't accord well with the rational-choice model, but might accord well with a self-control model.

The most important advantage of our self-control model, however, is that it permits more accurate *welfare* conclusions. Welfare conclusions are central to many economic analyses, but such conclusions are usually drawn under the maintained assumption that people always do what's best for themselves. In realms such as addiction where self-control problems and other errors seem likely to play an important role, such conclusions may be very misleading. Many observers — including, we suspect, many economists — believe that people develop and maintain addictions against their long-run best interests, and cause themselves severe harm in the process. If so, it is important to understand how and why people are hurting themselves, so that policies can be enacted to help people not to hurt themselves. By *a priori* assuming that people always act in their own self-interest, the rational-choice model precludes itself from answering these questions. We believe the economic method will prove useful in answering such questions, and we hope our self-control

[42] Indeed, every study with which we are familiar that has explicitly compared the empirical fit of different discount functions supports present-biased preferences over time-consistent preferences (and also over "future-biased preferences").

[43] As our analysis indicates, for most simple *qualitative* comparative statics, such as the effects of price changes, the rational-choice model and our model make the same predictions — because both models make the intuitively correct predictions. But we predict that for more complicated comparative statics our model may make sounder qualitative predictions as well.

model represents a useful step in this direction.[44]

The most basic lessons from our analysis are that self-control problems are a source of over-consumption of addictive products, and that awareness of self-control problems can mitigate or exacerbate this over-consumption. As we have emphasized throughout, however, our main concern is with quantitative results about whether self-control problems are a plausible source of severely harmful addictions. Our analysis suggests two possible ways in which self-control problems might cause severe harm. First, to the extent that a person is sophisticated, he may suffer severe harm due to feelings of inevitability. Second, to the extent that a person is naive, he may suffer severe harm due to procrastination in quitting an established addiction. But in real-world environments, lifelong feelings of inevitability seem implausible, while at the same time non-stationarities in the temptation to consume seem prevalent. Hence, our analysis suggests that for realistic environments, self-control problems are plausible source of harmful addictions only in conjunction with at least some degree of naivete.

While the most likely source harm from naivete is simple over-consumption, we conclude with one final example that indicates a second way in which naivete can cause harm in the face of non-stationarities:

Example 8: Suppose $\delta = .9$, $\beta = .6$, and $\gamma = .5$. Let $f(k) = -4k$, $g(k) = -24k$, and suppose $\bar{x}_t = 17$ for t odd and $\bar{x}_t = -16$ for t even. Then starting from $k_1 = 2/3$:

(1) TCs never hit,

(2) Naifs hit in odd periods but refrain in even periods, and

(3) Sophisticates hit always.

In Example 8, the exogenous temptation to hit fluctuates between a very high level and a very low level. While sophisticates consume more than naifs in this example, they are in fact suffering less harm.[45] Both are consuming more than is optimal, but the harm from consumption is very much not monotonic in consumption — if a person simply cannot sufficiently control himself, he may in fact be better off succumbing fully to his addiction rather than trying to eliminate it. Misguided and unpleasant attempts to quit addictions, followed by relapse, may represent another significant

[44] The reader might worry that we are proposing an overly paternalistic approach to policy. As we discuss elsewhere (see in particular O'Donoghue and Rabin (1999*c*)), we believe one should approach policy with a "cautious paternalism" wherein we look for policies that can be beneficial for people who make errors while having very little effect for people who are fully rational.

[45] It is easy to show that $U_1(k_1, \alpha^s) > U_1(k_1, \alpha^n)$.

problem for naifs.

Whether it be the unpleasantness of failed attempts to quit or the more fundamental problem of over-consumption, we share many non-economists conjecture that self-control problems are a major facet of cigarette, alcohol, and other forms of addiction. If economists want to contribute to the policy debate over how to deal with addictions, we need to develop a systematic approach to analyzing self-control problems and other errors rather than assume them away. We hope our analysis will prove useful in this regard.

Appendix: Proofs

Proof of Lemma 1: For use in this and other proofs, we define some additional notation. Define $\mathcal{A}^t \equiv \{0,1\}^{T-t+1}$, where $\mathbf{a}^t \equiv (a_t, a_{t+1}, ..., a_T) \in \mathcal{A}^t$ designates a behavior path beginning from period t. Define $V_t(k_t, \mathbf{a}^t)$ to be long-run continuation utility from following behavior path $\mathbf{a}^t$ given period-t addiction level k_t. Define $K_\tau(k_t, \mathbf{a}^t) \equiv \gamma^{\tau-t} k_t + \sum_{i=t}^{\tau-1} \gamma^{\tau-i-1} a_i$, which is the person's addiction level in period $\tau \geq t$ conditional on following $\mathbf{a}^t$ starting from addiction level k_t. Then

$$V_t(k_t, \mathbf{a}^t) = \sum_{\tau=t}^{T} \delta^{\tau-t} \left[a_\tau \left(x_o + f_\tau \left(K_\tau(k_t, \mathbf{a}^t) \right) \right) + (1 - a_\tau) \left(y_o + g_\tau \left(K_\tau(k_t, \mathbf{a}^t) \right) \right) \right].$$

By assumption f_t and g_t are weakly convex, and K_τ is increasing and linear in k_t, and therefore V_t is weakly convex in k_t. We assume throughout that $\max_{\mathbf{a} \in \mathcal{A}^t} V^t(k, \mathbf{a})$ exists for all k and t (see footnote 14).

(1) Since TCs are time-consistent, any perception-perfect strategy α^{tc} must satisfy for all t

$$U_t(k, \alpha^{tc}) = \max_{\mathbf{a} \in \mathcal{A}^t} V^t(k, \mathbf{a}).$$

To prove uniqueness, suppose that α^{tc} and $\hat{\alpha}^{tc}$ are both perception-perfect strategies for TCs. Then $U_t(k, \alpha^{tc}) = U_t(k, \hat{\alpha}^{tc}) = \max_{\mathbf{a} \in \mathcal{A}^t} V^t(k, \mathbf{a})$ for all k and t. By Definition 3, $\alpha^{tc}(k,t) = 1$ if and only if $h_t(k) \geq \delta \left[U_{t+1}(\gamma k, \alpha^{tc}) - U_{t+1}(\gamma k + 1, \alpha^{tc}) \right]$, and $\hat{\alpha}^{tc}(k,t) = 1$ if and only if $h_t(k) \geq \delta \left[U_{t+1}\left(\gamma k, \hat{\alpha}^{tc}\right) - U_{t+1}\left(\gamma k + 1, \hat{\alpha}^{tc}\right) \right]$. But then $U_t(k, \alpha^{tc}) = U_t(k, \hat{\alpha}^{tc})$ for all k and t implies $\alpha^{tc}(k,t) = \hat{\alpha}^{tc}(k,t)$ for all k and t — that is, α^{tc} and $\hat{\alpha}^{tc}$ must be the same strategy.

For all t, $U_t(k, \alpha^{tc})$ is the upper envelope of the set of weakly convex functions $V_t(k, \mathbf{a}^t)$, $\mathbf{a}^t \in \mathcal{A}^t$, and is therefore weakly convex in k. Hence, for all t, $\left[U_{t+1}(\gamma k, \alpha^{tc}) - U_{t+1}(\gamma k + 1, \alpha^{tc}) \right]$ is weakly decreasing in k. By Definition 3, $\alpha^{tc}(k,t) = 1$ if and only if $h_t(k) \geq$
$\delta \left[U_{t+1}(\gamma k, \alpha^{tc}) - U_{t+1}(\gamma k + 1, \alpha^{tc}) \right]$. Since $h_t(k)$ is increasing in k for all t, it follows that for all t there exists $\bar{k}_t^{tc}$ such that $\alpha^{tc}(k,t) = 1$ if and only if $k \geq \bar{k}_t^{tc}$.

(2) By Definition 4, $\alpha^n(k,t) = 1$ if and only if $h_t(k) \geq \beta\delta \left[U_{t+1}(\gamma k, \alpha^{tc}) - U_{t+1}(\gamma k + 1, \alpha^{tc}) \right]$. Given α^{tc} is unique, $\alpha^n(k,t)$ is uniquely defined for all k and t. Because for all t,
$U_{t+1}(\gamma k, \alpha^{tc}) - U_{t+1}(\gamma k + 1, \alpha^{tc})$ is weakly decreasing in k and $h_t(k)$ is increasing in k, it follows that for all t there exists $\bar{k}_t^n$ such that $\alpha^n(k,t) = 1$ if and only if $k \geq \bar{k}_t^n$.

(3) Since f_t and g_t are both decreasing in k for all t, $V^t(k, \mathbf{a})$ is decreasing in k for all t and $\mathbf{a} \in \mathcal{A}^t$, and therefore $U_t(k, \alpha^{tc})$ is decreasing in k for all t. This implies $U_{t+1}(\gamma k, \alpha^{tc}) - U_{t+1}(\gamma k + 1, \alpha^{tc}) > 0$ for all k and t. Since $\alpha^{tc}(k,t) = 1$ if and only if $h_t(k) \geq$

$\delta\left[U_{t+1}\left(\gamma k,\alpha^{tc}\right)-U_{t+1}\left(\gamma k+1,\alpha^{tc}\right)\right]$, whereas $\alpha^{n}(k,t)=1$ if and only if $h_t(k)\geq$ $\beta\delta\left[U_{t+1}\left(\gamma k,\alpha^{tc}\right)-U_{t+1}\left(\gamma k+1,\alpha^{tc}\right)\right]$, it follows that $\alpha^{tc}(k,t)\leq\alpha^{n}(k,t)$ for all k and t, which in turn implies $\bar{k}_t^{tc}\geq\bar{k}_t^{n}$ for all t.

QED

Proof of Lemma 2: Define $\mathbf{a}^{t+1}\equiv(a_{t+1},...,a_T)$ and $\mathbf{a}^{t+1\prime}\equiv(a'_{t+1},...,a'_T)$. Define $k_{\tau L}=K_\tau(\gamma k_t,\mathbf{a}^{t+1})$, $k_{\tau H}=K_\tau(\gamma k_t+1,\mathbf{a}^{t+1})$, $k'_{\tau L}=K_\tau(\gamma k_t,\mathbf{a}^{t+1\prime})$, and $k'_{\tau H}=K_\tau(\gamma k_t+1,\mathbf{a}^{t+1\prime})$. Note that for all τ, $k_{\tau L}-k_{\tau H}=k'_{\tau L}-k'_{\tau H}=\gamma^{\tau-t-1}$. Moreover, $a_\tau\geq a'_\tau$ for all τ implies $k_{\tau L}\geq k'_{\tau L}$ and $k_{\tau H}\geq k'_{\tau H}$ for all τ. Let $I(E)$ be an indicator function that takes a value of 1 if E is true and 0 otherwise. Then for α and α' as described in the premise,

$[U_{t+1}(\gamma k_t,\alpha)-U_{t+1}(\gamma k_t+1,\alpha)]-[U_{t+1}(\gamma k_t,\alpha')-U_{t+1}(\gamma k_t+1,\alpha')]=$

$\sum_{\tau=t}^{T}I(a_\tau=a'_\tau=1)\delta^{\tau-t}\left[(f_\tau(k_{\tau L})-f_\tau(k_{\tau H}))-(f_\tau(k'_{\tau L})-f_\tau(k'_{\tau H}))\right]$

$+\sum_{\tau=t}^{T}I(a_\tau=a'_\tau=0)\delta^{\tau-t}\left[(g_\tau(k_{\tau L})-g_\tau(k_{\tau H}))-(g_\tau(k'_{\tau L})-g_\tau(k'_{\tau H}))\right]$

$+\sum_{\tau=t}^{T}I(a_\tau>a'_\tau)\delta^{\tau-t}\left[(f_\tau(k_{\tau L})-f_\tau(k_{\tau H}))-(g_\tau(k'_{\tau L})-g_\tau(k'_{\tau H}))\right]$.

Given $k_{\tau L}-k_{\tau H}=k'_{\tau L}-k'_{\tau H}$ and $k_{\tau L}\geq k'_{\tau L}$, f_τ weakly convex implies $(f_\tau(k_{\tau L})-f_\tau(k_{\tau H}))-(f_\tau(k'_{\tau L})-f_\tau(k'_{\tau H}))\leq 0$ for all τ, and g_τ weakly convex implies $(g_\tau(k_{\tau L})-g_\tau(k_{\tau H}))-(g_\tau(k'_{\tau L})-g_\tau(k'_{\tau H}))\leq 0$ for all τ. Finally, $(f_\tau(k_{\tau L})-f_\tau(k_{\tau H}))-(g_\tau(k'_{\tau L})-g_\tau(k'_{\tau H}))=$ $(f_\tau(k_{\tau L})-f_\tau(k_{\tau H}))-(f_\tau(k'_{\tau L})-f_\tau(k'_{\tau H}))+h(k'_{\tau L})-h(k'_{\tau H})\leq 0$ for all τ because $k'_{\tau L}\leq k'_{\tau H}$ and h is increasing. Hence, $[U_{t+1}(\gamma k_t,\alpha)-U_{t+1}(\gamma k_t+1,\alpha)]-[U_{t+1}(\gamma k_t,\alpha')-U_{t+1}(\gamma k_t+1,\alpha')]\leq 0$, and the result follows.

QED

Proof of Lemma 3: $T=\infty$ implies that $\mathcal{A}^t=\{0,1\}^\infty$ for all t, and then stationary preferences imply that $V^t(k_t,\mathbf{a}^t)$ is independent of t, and therefore $U_t(k,\alpha^{tc})=\max_{\mathbf{a}\in\mathcal{A}^t}V^t(k,\mathbf{a})$ is independent of t. Stationary preferences also imply that $h_t(k)$ is independent of t. Because $\alpha^{tc}(k,t)=1$ if and only if $h_t(k)\geq\delta\left[U_{t+1}\left(\gamma k,\alpha^{tc}\right)-U_{t+1}\left(\gamma k+1,\alpha^{tc}\right)\right]$, it follows that $\alpha^{tc}(k,t)$ is independent of t, in which case Lemma 1(1) implies that TCs have a stationary cutoff $\bar{k}^{tc}$. Similarly, because $\alpha^{n}(k,t)=1$ if and only if $h_t(k)\geq\beta\delta\left[U_{t+1}\left(\gamma k,\alpha^{tc}\right)-U_{t+1}\left(\gamma k+1,\alpha^{tc}\right)\right]$, it follows that $\alpha^{n}(k,t)$ is independent of t, in which case Lemma 1(2) implies that naifs have a stationary cutoff $\bar{k}^{n}$.

QED

Proof of Lemma 4: When $T<\infty$, α^s is unique (since we assume a person hits when indifferent).

Our "limit-of-the-finite-horizon" reasoning involves solving for this strategy and asking what it looks like far from the end of the game. To make such arguments, two new pieces of notation will be useful:

Define $\tilde{\Delta}^H(k,\eta)$ to be the future cost from hitting for a person whose current addiction level is k who will hit no matter what in the η remaining periods. Formally,

$$\tilde{\Delta}^H(k,\eta) = \sum_{n=1}^{\eta} \delta^{n-1} \left[f\left(\gamma^n k + \sum_{m=1}^{n-1} \gamma^{m-1}\right) - f\left(\gamma^n k + \sum_{m=1}^{n} \gamma^{m-1}\right)\right] \geq 0.$$

$\tilde{\Delta}^H$ is weakly decreasing in k, is increasing in η, and $\Delta^H = \tilde{\Delta}^H(0,\infty)$.

Define $\mathbf{r}_\tau^t \equiv \{(a_t, a_{t+1}, ..., a_T) \in \mathcal{A}^t \mid a_{t'} = 1 \text{ if and only if } t' \geq \tau\}$. In words, $\mathbf{r}_\tau^t$ is the period-t behavior path that involves refraining until period τ and then hitting thereafter.

Suppose $\bar{x}_o \geq \beta\delta\Delta^H$. Because this implies $h(0) \geq 0$, $\alpha^s(k,T) = 1$ for all $k \geq 0$ — the person hits no matter what in period T. Given this, $\alpha^s(k,T-1) = 1$ if and only if $h_{T-1}(k) \geq \beta\delta\tilde{\Delta}^H(k,1)$. Given $h_{T-1}(k) \geq \bar{x}_o$ for all k and $\tilde{\Delta}^H(k,1) \leq \Delta^H$ for all k, $\bar{x}_o \geq \beta\delta\Delta^H$ implies $\alpha^s(k,T-1) = 1$ for all $k \geq 0$ — the person hits no matter what in period $T-1$. Iterating this logic, it is straightforward to derive that for any $T < \infty$, $\alpha^s(k,t) = 1$ for all k and t, in which case the corresponding infinite-horizon strategy involves $\alpha^s(k,t) = 1$ for all k and t.

Suppose $\bar{x}_o < \beta\delta\Delta^H$. Now there exists $\bar{\eta} \in \{0,1,...\}$ such that $\bar{x}_o < \beta\delta\tilde{\Delta}^H(0,\eta)$ if and only if $\eta \geq \bar{\eta}$. The logic above implies that for all $t > T - \bar{\eta}$, $\alpha^s(k,t) = 1$ for all k, and that for $\bar{\tau} \equiv T - \bar{\eta}$ there exists $k' > 0$ such that $\alpha^s(k,\bar{\tau}) = 0$ if and only if $k < k'$. It is straightforward (although tedious) to derive that for any $t < \bar{\tau}$ and $k < k'$, a sophisticate's desired behavior conditional on hitting from period $\bar{\tau}+1$ onward — that is, among the set of strategies $\mathcal{A}_*^t \equiv$ $\{(a_t, ...a_T) \in \mathcal{A}^t \mid a_\tau = 1 \text{ for } t > \bar{\tau}\}$ — is $\mathbf{r}_{\bar{\tau}+1}^t$. Hence, in period $\bar{\tau}-1$, a sophisticate with $k_{\bar{\tau}-1} < k'$ perceives that refraining now will lead to following his desired behavior path $\mathbf{r}_{\bar{\tau}+1}^{\bar{\tau}-1}$, and so $\alpha^s(k,\bar{\tau}-1) = 0$ for all $k < k'$. But this means that in period $\bar{\tau}-2$ a sophisticate with $k_{\bar{\tau}-2} < k'$ perceives that refraining now will lead to following his desired behavior path $\mathbf{r}_{\bar{\tau}+1}^{\bar{\tau}-2}$, and so $\alpha^s(k,\bar{\tau}-2) = 0$ for all $k < k'$. Iterating this logic, it follows that for all $t \leq \bar{\tau}$, $\alpha^s(k,t) = 0$ for all $k < k'$, in which case the corresponding infinite-horizon strategy involves for all t, $\alpha^s(k,t) = 0$ for all $k < k'$.

QED

Proof of Lemma 5: The value $k^*(\beta)$ is the k^* such that

$\frac{(1-\delta+\beta\delta)\bar{x}_o}{1-\delta} + f(k^*) + \beta\delta\sum_{n=1}^{\infty}\delta^{n-1}f\left(\gamma^n k^* + \sum_{m=0}^{n-1}\gamma^m\right) = g(k^*) + \beta\delta\sum_{n=1}^{\infty}\delta^{n-1}g(\gamma^n k^*)$,

and the value $\tilde{k}(\beta)$ is the $\tilde{k}$ such that

$\bar{x}_o + f\left(\tilde{k}\right) + \beta\delta \sum_{n=1}^{\infty} \delta^{n-1} g\left(\gamma^n \tilde{k} + \gamma^{n-1}\right) = g\left(\tilde{k}\right) + \beta\delta \sum_{n=1}^{\infty} \delta^{n-1} g\left(\gamma^n \tilde{k}\right)$.

(1) For all k and t, TCs follow their desired behavior path, which is either hit always or refrain always. Clearly $\alpha^{tc}(k,t) = 1$ if and only if their desired behavior path is hit always, which means $k \geq k^*(1)$.

(2) For all k and t, naifs attempt to follow their desired behavior path, which is either hit always, hit once, or refrain always. Hence, $\alpha^{n}(k,t) = 1$ if and only if their desired behavior path is either hit always or hit once, which means $k \geq \min\{k^*(\beta), \tilde{k}(\beta)\}$.

(3) That $\alpha^{s}(k,t) = 1$ for all k if $\bar{x}_o \geq \beta\delta\Delta^H$ follows from Lemma 4. Suppose $\bar{x}_o < \beta\delta\Delta^H$. Define k^{**} such that sophisticates' desired behavior path is hit always for all $k \geq k^{**}$, in which case clearly $\alpha^{s}(k,t) = 1$ if $k \geq k^{**}$. Consider $k \in [(k^{**}-1)/\gamma, k^{**})$. Because $\gamma k + 1 \geq k^{**}$, hitting with addition level k will lead a sophisticate to hit always. The best possible behavior path following restraint is never hitting (given that $k < k^{**}$). Hence, for any $k \in [(k^{**}-1)/\gamma, k^{**})$ such that $k \geq k^*(\beta)$, which means hitting always is preferred to never hitting, $\alpha^{s}(k,t) = 1$. Because we can iterate this logic, it follows that for any t, $\alpha^{s}(k,t) = 1$ if $k \geq k^*(\beta)$.

By Lemma 4, there exists $k' > 0$ such that for all t, $\alpha^{s}(k,t) = 0$ if $k < k'$. Consider $k \in [k', k'/\gamma)$. Because $\gamma k < k'$, refraining with addiction level k implies never hitting. Hence, for any $k \in [k', k'/\gamma)$ such that $k < \min\{k^*(\beta), \tilde{k}(\beta)\}$, which means never hitting is sophisticates' desired behavior path, $\alpha^{s}(k,t) = 0$. Because we can iterate this logic, it follows that for any t, $\alpha^{s}(k,t) = 0$ if $k < \min\{k^*(\beta), \tilde{k}(\beta)\}$.

Suppose $\gamma\tilde{k}(\beta) + 1 \geq k^*(\beta)$. In this case, one possibility is $\tilde{k}(\beta) \geq k^*(\beta)$, in which case it follows from above that $\alpha^{s}(k,t) = 1$ if and only if $k \geq k^*(\beta)$. The other possibility is $\tilde{k}(\beta) < k^*(\beta)$. But then for any $k \in [\tilde{k}(\beta), k^*(\beta))$, hitting with addiction level k implies hitting always (because $\gamma k + 1 \geq k^*(\beta)$), and an iteration logic similar to that in the previous paragraph yields $\alpha^{s}(k,t) = 1$ if and only if $k \geq k^*(\beta)$.

Finally suppose $\gamma\tilde{k}(\beta) + 1 < k^*(\beta)$, which implies $\tilde{k}(\beta) < k^*(\beta)$. In this case, our results above imply $\alpha^{s}(k,t) = 0$ if $k < \tilde{k}(\beta)$ and $\alpha^{s}(k,t) = 1$ if $k \geq k^*(\beta)$.

QED

Proof of Proposition 1: (1) $\alpha^{tc}(k,t) \leq \alpha^{n}(k,t)$ for all k and t is established by Lemma 1; and $\alpha^{n}(k,t) \leq \alpha^{s}(k,t)$ for all k and t follows trivially from Lemma 4, which establishes that $\bar{x}_o \geq \beta\delta\Delta^H$ implies $\alpha^{s}(k,t) = 1$ for all k and t.

(2) $\alpha^s(k,t) \leq \alpha^n(k,t)$ follows from Lemma 5, which establishes that $\alpha^n(k,t) = 1$ if $k \geq \min\{k^*(\beta), \tilde{k}(\beta)\}$ whereas $\alpha^s(k,t) = 1$ only if $k \geq \min\{k^*(\beta), \tilde{k}(\beta)\}$. $\alpha^{tc}(k,t) \leq \alpha^s(k,t)$ also follows from Lemma 5, which establishes $\alpha^s(k,t) = 1$ if $k \geq k^*(\beta)$ whereas $\alpha^{tc}(k,t) = 1$ only if $k \geq k^*(1) \geq k^*(\beta)$.

QED

Proof of Proposition 2: Define $\bar{k}$ such that

$$f(\bar{k}) - g(\bar{k}) = \beta\delta\left[\Delta^R\left(\bar{k}\right) - \Delta^H\right]$$

where $\Delta^R(k) = \sum_{n=1}^{\infty} \delta^{n-1}\left[g\left(\gamma^n k\right) - g\left(\gamma^n k + \gamma^{n-1}\right)\right]$. Because $f(k) - g(k)$ is increasing in k, and because $\beta\delta\left[\Delta^R(k) - \Delta^H\right]$ is weakly decreasing in k (since g is weakly convex), there exists a unique such $\bar{k}$, and moreover $f(k) - g(k) < \beta\delta\left[\Delta^R(k) - \Delta^H\right]$ for $k < \bar{k}$ and $f(k) - g(k) > \beta\delta\left[\Delta^R(k) - \Delta^H\right]$ for $k > \bar{k}$. Because $f(0) - g(0) = 0 < \beta\delta\left[\Delta^R(0) - \Delta^H\right]$ (the inequality follows from Lemma 2), $\bar{k} > 0$. Because $\delta = 1$ implies $\beta\delta\left[\Delta^R(k^{\max}) - \Delta^H\right] = \beta\left[f(k^{\max}) - g(k^{\max})\right]$, and because $\beta\delta\left[\Delta^R(k) - \Delta^H\right]$ is increasing in δ (whenever it's positive), $\bar{k} < k^{\max}$.

(1) Suppose $k_1 \leq \bar{k}$. By Lemma 5, naifs hit always only if $\bar{x}_o$ such that $k_1 \geq \min\{k^*(\beta), \tilde{k}(\beta)\}$. If $k_1 \geq k^*(\beta)$, then sophisticates hit always. If $k_1 \geq \tilde{k}(\beta)$, then $\bar{x}_o + f(k_1) - g(k_1) \geq \beta\delta\Delta^R(k_1)$. But since $k_1 \leq \bar{k}$ implies $f(k_1) - g(k_1) \leq \beta\delta\Delta^R(k_1) - \beta\delta\Delta^H$, $k_1 \geq \tilde{k}(\beta)$ implies $\bar{x}_o \geq \beta\delta\Delta^H$ and therefore sophisticates hit always. The result follows.

(2) Suppose $k_1 \geq \bar{k}$. By Lemma 5, naifs never hit only if $\bar{x}_o$ such that $k_1 < \min\{k^*(\beta), \tilde{k}(\beta)\}$, which requires $k_1 < \tilde{k}(\beta)$. If $k_1 < \tilde{k}(\beta)$ then $\bar{x}_o + f(k_1) - g(k_1) < \beta\delta\Delta^R(k_1)$. And since $k_1 \geq \bar{k}$ implies $f(k_1) - g(k_1) \geq \beta\delta\Delta^R(k_1) - \beta\delta\Delta^H$, $k_1 < \tilde{k}(\beta)$ implies $\bar{x}_o < \beta\delta\Delta^H$ and therefore sophisticates refrain whenever naifs refrain. The result follows.

QED

Proof of Proposition 3: It is clear that both behavior and welfare losses depend only on $\bar{x}_o \equiv x_o - y_o$ and not on the specific values of x_o and y_o. For notational simplicity, therefore, the proofs of Propositions 3 and 4 shall assume $y_o = 0$ and $x_o = \bar{x}_o$. Define $V^{NH}(\bar{x}_o, k_1)$, $V^{HA}(\bar{x}_o, k_1)$, and $V^{H1}(\bar{x}_o, k_1)$ to be the long-run utilities from never hitting, hitting always, and hitting once, respectively. Similarly, define $\tilde{V}^{NH}(\bar{x}_o, k_1)$, $\tilde{V}^{HA}(\bar{x}_o, k_1)$, and $\tilde{V}^{H1}(\bar{x}_o, k_1)$ to be the short-run utilities from these behavior paths. Given $f(k) = -\rho k$, $g(k) = -\phi\rho k$, and $\Delta^H = \frac{\rho}{1-\delta\gamma}$, it is straightforward to derive:

$V^{NH}(\bar{x}_o, k_1) = -\phi\Delta^H k_1$ and $\tilde{V}^{NH}(\bar{x}_o, k_1) = -\phi\rho k_1 - \beta\delta\phi\Delta^H\gamma k_1$

$V^{HA}(\bar{x}_o, k_1) = \frac{\bar{x}_o}{1-\delta} - \frac{\delta\Delta^H}{1-\delta} - \Delta^H k_1$ and $\tilde{V}^{HA}(\bar{x}_o, k_1) = \bar{x}_o - \rho k_1 + \frac{\beta\delta\bar{x}_o}{1-\delta} - \frac{\beta\delta\Delta^H}{1-\delta} - \beta\delta\Delta\gamma k_1$.

$V^{H1}(\bar{x}_o, k_1) = \bar{x}_o - \rho k_1 - \delta\phi\Delta^H(\gamma k_1 + 1)$ and $\tilde{V}^{H1}(\bar{x}_o, k_1) = \bar{x}_o - \rho k_1 - \beta\delta\phi\Delta^H(\gamma k_1 + 1)$.

(1) TCs either never hit or hit always; committers either never hit, hit always, or hit once. Committers suffer welfare losses only if they hit always or hit once when TCs never hit. Because $\frac{\partial V^{HA}}{\partial \bar{x}_o} > \frac{\partial V^{H1}}{\partial \bar{x}_o} > \frac{\partial V^{NH}}{\partial \bar{x}_o}$, welfare losses are maximized at either the minimum $\bar{x}_o$ such that committers hit always or the minimum $\bar{x}_o$ such that committers hit once.

Define $x^{NH,H1}$ such that $\tilde{V}^{NH}(x^{NH,H1}, 0) = \tilde{V}^{H1}(x^{NH,H1}, 0)$, define $x^{NH,HA}$ such that $\tilde{V}^{NH}(x^{NH,HA}, 0) = \tilde{V}^{HA}(x^{NH,HA}, 0)$, and define $x^{H1,HA}$ such that $\tilde{V}^{H1}(x^{H1,HA}, 0) = \tilde{V}^{HA}(x^{H1,HA}, 0)$. Algebra reveals $x^{NH,HA} = \frac{\beta\delta\Delta^H}{1-\delta+\beta\delta}$ and $x^{NH,H1} = \beta\delta\phi\Delta^H$, and therefore $x^{NH,H1} \leq x^{NH,HA}$ if and only if $\phi \leq \frac{1}{1-\delta+\beta\delta}$.

Suppose $\phi \leq \frac{1}{1-\delta+\beta\delta}$. Because $\frac{\partial \tilde{V}^{HA}}{\partial \bar{x}_o} > \frac{\partial \tilde{V}^{H1}}{\partial \bar{x}_o} > \frac{\partial \tilde{V}^{NH}}{\partial \bar{x}_o}$, it follows that $x^{H1,HA} \geq x^{NH,H1}$, and therefore committers never hit for $\bar{x}_o < x^{NH,H1}$, hit once for $\bar{x}_o \in [x^{NH,H1}, x^{H1,HA})$, and hit always for $\bar{x}_o \geq x^{H1,HA}$. Hence, if $\phi \leq \frac{1}{1-\delta+\beta\delta}$ welfare losses are maximized at $x^{NH,H1}$ (because $x^{H1,HA}$ is the $\bar{x}_o$ at which TCs are indifferent between hitting once and hitting always), and so $\max_{\bar{x}_o \in \mathbf{R}} \left[WL(0, \alpha^{\text{commit}})\right] = V^{NH}(x^{NH,H1}, 0) - V^{H1}(x^{NH,H1}, 0) = \delta(1-\beta)\phi\Delta^H$.

If $\phi \geq \frac{1}{1-\delta+\beta\delta}$ then $x^{NH,H1} \geq x^{NH,HA}$, in which case $\frac{\partial \tilde{V}^{HA}}{\partial \bar{x}_o} > \frac{\partial \tilde{V}^{H1}}{\partial \bar{x}_o} > \frac{\partial \tilde{V}^{NH}}{\partial \bar{x}_o}$ implies that committers never hit for $\bar{x}_o < x^{NH,HA}$ and hit always for $\bar{x}_o \geq x^{NH,HA}$. Hence, if $\phi \geq \frac{1}{1-\delta+\beta\delta}$ then welfare losses are maximized at $x^{NH,HA}$, and so $\max_{\bar{x}_o \in \mathbf{R}} \left[WL(0, \alpha^{\text{commit}})\right] = V^{NH}(x^{NH,HA}, 0) - V^{HA}(x^{NH,HA}, 0) = \frac{\delta(1-\beta)}{1-\delta+\beta\delta}\Delta^H$.

(2) TCs and sophisticates both either never hit or hit always, and so sophisticates suffer welfare losses only if they hit always when TCs never hit. Because $\frac{\partial V^{HA}}{\partial \bar{x}_o} > \frac{\partial V^{NH}}{\partial \bar{x}_o}$, welfare losses are maximized at the minimum $\bar{x}_o$ such that sophisticates hit always. Lemma 4 implies sophisticates hit always if and only if $\bar{x}_o \geq \beta\delta\Delta^H$. Hence, for any $\phi > 1$ welfare losses are maximized at $\bar{x}_o = \beta\delta\Delta^H$, and so $\max_{\bar{x}_o \in \mathbf{R}} \left[WL(0, \alpha^s)\right] = V^{NH}(\beta\delta\Delta^H, 0) - V^{HA}(\beta\delta\Delta^H, 0) = \frac{\delta(1-\beta)}{1-\delta}\Delta^H$.

(3) Like sophisticates, naifs suffer welfare losses only if they hit always when TCs never hit, and so welfare losses are maximized at the minimum $\bar{x}_o$ such that naifs hit always. Using the notation from the proof of part 1, naifs hit always if $\bar{x}_o \geq \min\{x^{NH,HA}, x^{NH,H1}\}$.

As above, if $\phi \leq \frac{1}{1-\delta+\beta\delta}$ then $x^{NH,H1} \leq x^{NH,HA}$, and therefore welfare losses are maximized at $x^{NH,H1}$. Hence, $\max_{\bar{x}_o \in \mathbf{R}} \left[WL(0, \alpha^n)\right] = V^{NH}(x^{NH,H1}, 0) - V^{HA}(x^{NH,H1}, 0) = \frac{\delta(1-\beta\phi)}{1-\delta}\Delta^H$.

If $\phi \geq \frac{1}{1-\delta+\beta\delta}$ then $x^{NH,H1} \geq x^{NH,HA}$, and therefore welfare losses are maximized at $x^{NH,HA}$.

Hence, $\max_{\bar{x}_o \in \mathbf{R}} [WL(0, \alpha^n)] = V^{NH}(x^{NH,HA}, 0) - V^{HA}(x^{NH,HA}, 0) = \frac{\delta(1-\beta)}{1-\delta+\beta\delta}\Delta^H$.

QED

Proof of Proposition 4: (1) For any k_1, committers suffer welfare losses only if they hit always or hit once when TCs never hit, and welfare losses are maximized at either the minimum $\bar{x}_o$ such that committers hit always or the minimum $\bar{x}_o$ such that committers hit once.

Define $x^{NH,H1}$ such that $\tilde{V}^{NH}(x^{NH,H1}, k_1) = \tilde{V}^{H1}(x^{NH,H1}, k_1)$, define $x^{NH,HA}$ such that $\tilde{V}^{NH}(x^{NH,HA}, k_1) = \tilde{V}^{HA}(x^{NH,HA}, k_1)$, and define $x^{H1,HA}$ such that $\tilde{V}^{H1}(x^{H1,HA}, k_1) = \tilde{V}^{HA}(x^{H1,HA}, k_1)$. Algebra reveals $x^{NH,HA} = \frac{\beta\delta\Delta^H}{1-\delta+\beta\delta} - \frac{(1-\delta)(1-\delta\gamma+\beta\delta\gamma)}{1-\delta+\beta\delta}(\phi - 1)\Delta^H k_1$ and $x^{NH,H1} = \beta\delta\phi\Delta^H - (1-\delta\gamma)(\phi-1)\Delta^H k_1$, and therefore $x^{NH,H1} \leq x^{NH,HA}$ if and only if $k_1 \geq \frac{(1-\delta+\beta\delta)\phi-1}{\phi-1}k^{\max} \equiv k^*$ (recall that $k^{\max} = \frac{1}{1-\gamma}$). Note that $\phi > \frac{1}{1-\delta+\beta\delta}$ implies $k^* \in (0, (1-\delta+\beta\delta)k^{\max})$.

Suppose $k_1 \geq k^*$. Because $\frac{\partial\tilde{V}^{HA}}{\partial\bar{x}_o} > \frac{\partial\tilde{V}^{H1}}{\partial\bar{x}_o} > \frac{\partial\tilde{V}^{NH}}{\partial\bar{x}_o}$, it follows that $x^{H1,HA} \geq x^{NH,H1}$, and therefore committers never hit for $\bar{x}_o < x^{NH,H1}$, hit once for $\bar{x}_o \in [x^{NH,H1}, x^{H1,HA})$, and hit always for $\bar{x}_o \geq x^{H1,HA}$. Hence, if $k_1 \geq k^*$ then welfare losses are maximized at $x^{NH,H1}$, and so $\max_{\bar{x}_o \in \mathbf{R}} \left[WL(k_1, \alpha^{\text{commit}})\right] = V^{NH}(x^{NH,H1}, k_1) - V^{H1}(x^{NH,H1}, k_1) = \delta(1-\beta)\phi\Delta^H$.

If $k_1 \leq k^*$ then $x^{NH,H1} \geq x^{NH,HA}$, in which case $\frac{\partial\tilde{V}^{HA}}{\partial\bar{x}_o} > \frac{\partial\tilde{V}^{H1}}{\partial\bar{x}_o} > \frac{\partial\tilde{V}^{NH}}{\partial\bar{x}_o}$ implies that committers never hit for $\bar{x}_o < x^{NH,HA}$ and hit always for $\bar{x}_o \geq x^{NH,HA}$. Hence, if $k_1 \leq k^*$ then welfare losses are maximized at $x^{NH,HA}$, and so $\max_{\bar{x}_o \in \mathbf{R}} \left[WL(k_1, \alpha^{\text{commit}})\right] = V^{NH}(x^{NH,HA}, k_1) - V^{HA}(x^{NH,HA}, k_1) = \frac{\delta(1-\beta)}{1-\delta+\beta\delta}\Delta^H \left[1 + \frac{k_1}{k^{\max}}(\phi - 1)\right]$.

(2) Naifs suffer welfare losses only if they hit always when TCs never hit, and so welfare losses are maximized at the minimum $\bar{x}_o$ such that naifs hit always. Using the notation from the proof of part 1, naifs hit always if $\bar{x}_o \geq \min\{x^{NH,HA}, x^{NH,H1}\}$.

As above, if $k_1 \geq k^*$ then $x^{NH,H1} \leq x^{NH,HA}$, and therefore welfare losses are maximized at $x^{NH,H1}$. Hence, $\max_{\bar{x}_o \in \mathbf{R}} [WL(k_1, \alpha^n)] = V^{NH}(x^{NH,H1}, k_1) - V^{HA}(x^{NH,H1}, k_1) = \frac{\delta(1-\beta)\phi}{1-\delta}\Delta^H - \left(1 - \frac{k_1}{k^{\max}}\right)\frac{\delta(\phi-1)}{1-\delta}$.

If $k_1 \leq k^*$ then $x^{NH,H1} \geq x^{NH,HA}$, and therefore welfare losses are maximized at $x^{NH,HA}$. Hence, $\max_{\bar{x}_o \in \mathbf{R}} [WL(k_1, \alpha^n)] = V^{NH}(x^{NH,HA}, k_1) - V^{HA}(x^{NH,HA}, k_1) = \frac{\delta(1-\beta)}{1-\delta+\beta\delta}\Delta^H \left[1 + \frac{k_1}{k^{\max}}(\phi - 1)\right]$.

QED

Proof of Lemma 6: (1) We first prove that for any behavior paths $\mathbf{a} \equiv (a_1, a_2, ...)$ and $\mathbf{a}' \equiv$

$(a_1', a_2', ...)$ with $a_n \geq a_n'$ for all n, $V_t(k, \mathbf{a}) - V_t(k, \mathbf{a}') \geq V_\tau(k, \mathbf{a}) - V_\tau(k, \mathbf{a}')$ for any $t < \tau$. Given f and g are independent of t, $[V_t(k, \mathbf{a}) - V_t(k, \mathbf{a}')] - [V_\tau(k, \mathbf{a}) - V_\tau(k, \mathbf{a}')] =$
$\sum_{n=1}^{\infty} \delta^{n-1} I(a_n > a_n') \left[(x_{t+n} - y_{t+n}) - (x_{\tau+n} - y_{\tau+n})\right]$, where I is an indicator function as in the proof of Lemma 2. Given youthful instantaneous utilities, $t < \tau$ implies $x_{t+n} - y_{t+n} \geq x_{\tau+n} - y_{\tau+n}$ for all n, and the result follows.

Suppose that $\bar{k}_t^{tc} \leq \bar{k}_{t+1}^{tc}$ for all $t \geq \tau$. Letting $\tilde{\mathbf{a}}$ be the optimal behavior path for a person in period τ with addiction level $k_\tau = \bar{k}_\tau^{tc}$, which must involve hitting in period τ, and defining $\mathbf{r} \equiv (0, 0, ...)$, we must have $V_\tau(\bar{k}_\tau^{tc}, \tilde{\mathbf{a}}) \geq V_\tau(\bar{k}_\tau^{tc}, \mathbf{r})$. Now consider a person in period $\tau - 1$ with addiction level $k_{\tau-1} = \bar{k}_\tau^{tc}$. Given the premise that $\bar{k}_t^{tc} \leq \bar{k}_{t+1}^{tc}$ for all $t \geq \tau$, if this person refrains then he will refrain forever after. Given our assumption that people hit when indifferent, this person can therefore refrain only if $V_{\tau-1}(\bar{k}_\tau^{tc}, \mathbf{r}) > V_{\tau-1}(\bar{k}_\tau^{tc}, \mathbf{a})$ for all $\mathbf{a} \in \{0, 1\}^\infty$, and in particular only if $V_{\tau-1}(\bar{k}_\tau^{tc}, \mathbf{r}) > V_{\tau-1}(\bar{k}_\tau^{tc}, \tilde{\mathbf{a}})$. But our result in the previous paragraph implies that if $V_\tau(\bar{k}_\tau^{tc}, \tilde{\mathbf{a}}) \geq V_\tau(\bar{k}_\tau^{tc}, \mathbf{r})$ then $V_{\tau-1}(\bar{k}_\tau^{tc}, \tilde{\mathbf{a}}) \geq V_{\tau-1}(\bar{k}_\tau^{tc}, \mathbf{r})$. Hence, this person must hit, and therefore $\bar{k}_{\tau-1}^{tc} \leq \bar{k}_\tau^{tc}$.

We have thus established that if $\bar{k}_t^{tc} \leq \bar{k}_{t+1}^{tc}$ for all $t \geq \tau$, then $\bar{k}_t^{tc} \leq \bar{k}_{t+1}^{tc}$ for all $t \geq \tau - 1$. Lemma 3 implies that $\bar{k}_t^{tc} = \bar{k}_{t+1}^{tc}$ for all $t \geq M$, and the result follows.

(2) The proof is almost identical to that for TCs, and so is omitted.

(3) Note that if for all $t > \tau$ $\alpha^s(k, t) = 1$ for all k, then $U_{\tau+1}(\gamma k, \alpha^s) - U_{\tau+1}(\gamma k + 1, \alpha^s) = \tilde{\Delta}^H(k, \infty)$, where $\tilde{\Delta}^H$ is defined in the proof of Lemma 4 and is independent of τ. By Lemma 4, $\bar{x}_M \geq \beta\delta\Delta^H$ implies that for all $t \geq M$ $\alpha^s(k, t) = 1$ for all k, which in turn requires $h_M(k) \geq \beta\delta\tilde{\Delta}^H(k, \infty)$ for all k. Then $h_{M-1}(k) \geq h_M(k)$ for all k implies $h_{M-1}(k) \geq \beta\delta\tilde{\Delta}^H(k, \infty)$ for all k, and therefore $\alpha^s(k, M-1) = 1$ for all k. Iterating this logic, it follows that $\alpha^s(k, t) = 1$ for all k and t.

That $\bar{x}_M < \beta\delta\Delta^H$ implies there exists $k' > 0$ such that for all $t \geq M$ $\alpha^s(k, t) = 0$ for all $k < k'$ follows directly from Lemma 4.

QED

Proof of Proposition 5: (1) $\alpha^n(k, t) \leq \alpha^s(k, t)$ for all k and t follows trivially from Lemma 6, which establishes that $\bar{x}_M \geq \beta\delta\Delta^H$ implies $\alpha^s(k, t) = 1$ for all k and t.

(2) We first establish that if for all $t > \tau$ $\alpha^s(k, t) \leq \alpha^n(k, t)$ for all k, then $\alpha^s(k, \tau) \leq \alpha^n(k, \tau)$ for all k. If $\alpha^n(k, \tau) = 1$, then clearly $\alpha^s(k, \tau) \leq \alpha^n(k, \tau)$. Suppose instead that $\alpha^n(k, \tau) = 0$, in which case $h_\tau(k) < \beta\delta\left[U_{\tau+1}(\gamma k, \alpha^{tc}) - U_{\tau+1}(\gamma k + 1, \alpha^{tc})\right]$. By Lemma 6, if naifs refrain in pe-

riod τ then they will refrain forever after, which implies that if TCs refrain in period τ then they will refrain forever after. Moreover, the premise that for all $t > \tau$ $\alpha^s(k,t) \leq \alpha^n(k,t)$ for all k implies that if sophisticates refrain in period τ then they will refrain forever after, and so $U_{\tau+1}(\gamma k, \alpha^s) = U_{\tau+1}(\gamma k, \alpha^{tc})$. By revealed preference for TCs, $U_{\tau+1}(\gamma k+1, \alpha^{tc}) \geq U_{\tau+1}(\gamma k+1, \alpha^s)$, which implies $\beta\delta\left[U_{\tau+1}\left(\gamma k, \alpha^s\right) - U_{\tau+1}\left(\gamma k+1, \alpha^s\right)\right] \geq \beta\delta\left[U_{\tau+1}\left(\gamma k, \alpha^{tc}\right) - U_{\tau+1}\left(\gamma k+1, \alpha^{tc}\right)\right] > h_\tau(k)$ and therefore $\alpha^s(k,\tau) = 0$. The claim follows.

The result then follows from Proposition 1, which implies that if $\bar{x}_M < \beta\delta\Delta^H$ then $\alpha^s(k,t) \leq \alpha^n(k,t)$ for all k and $t \geq M$.

QED

Proof of Proposition 6: Let $w(\mathbf{a})$ and $\hat{w}(\mathbf{a})$ be the person's period-1 utility before and after the youthful rotation, respectively, from following behavior path $\mathbf{a}$ given initial addiction level $k_1 = 0$. Let $\mathbf{r} \equiv (0,0,...)$, $\mathbf{h} \equiv (1,1,...)$, and $\mathbf{h}^1 \equiv (1,0,0,...)$.

(1) For TCs, the definition of a youthful rotation implies $w(\mathbf{h}) = \hat{w}(\mathbf{h})$ and $w(\mathbf{r}) = \hat{w}(\mathbf{r})$. Because $\mathbf{a}^{tc} = \mathbf{h}$ only if $w(\mathbf{h}) \geq w(\mathbf{r})$ and $\widehat{\mathbf{a}}^{tc} = \mathbf{r}$ only if $\hat{w}(\mathbf{h}) < \hat{w}(\mathbf{r})$, $\mathbf{a}^{tc} = \mathbf{h}$ implies $\widehat{\mathbf{a}}^{tc} \neq \mathbf{r}$. Similarly, because $\mathbf{a}^{tc} = \mathbf{r}$ only if $w(\mathbf{r}) > w(\mathbf{h})$ and $\widehat{\mathbf{a}}^{tc} = \mathbf{h}$ only if $\hat{w}(\mathbf{r}) \leq \hat{w}(\mathbf{h})$, $\mathbf{a}^{tc} = \mathbf{r}$ implies $\widehat{\mathbf{a}}^{tc} \neq \mathbf{h}$.

(2) For naifs, the definition of a youthful rotation implies $w(\mathbf{h}) = \hat{w}(\mathbf{h})$, $w(\mathbf{r}) = \hat{w}(\mathbf{r})$, and $w(\mathbf{h}^1) \leq \hat{w}(\mathbf{h}^1)$. Lemma 5 implies $\mathbf{a}^n = \mathbf{h}$ only if $\min\{w(\mathbf{h}), w(\mathbf{h}^1)\} \geq w(\mathbf{r})$, and since $\min\{\hat{w}(\mathbf{h}), \hat{w}(\mathbf{h}^1)\} \geq \min\{w(\mathbf{h}), w(\mathbf{h}^1)\} \geq w(\mathbf{r}) = \hat{w}(\mathbf{r})$, it follows that $\widehat{\mathbf{a}}^n \neq \mathbf{r}$.

(3) $\mathbf{a}^s = \mathbf{r}$ implies $\mathbf{a}^n = \mathbf{r}$ by Proposition 2. $\mathbf{a}^s = \mathbf{r}$ also implies that $\bar{x}_o < \beta\delta\Delta^H$ by Lemma 4, and the definition of a youthful rotation then implies $\bar{x}_M < \beta\delta\Delta^H$. Proposition 5 then implies that if in addition $\widehat{\mathbf{a}}^s = \mathbf{h}$ then $\widehat{\mathbf{a}}^n = \mathbf{h}$.

QED

Proof of Proposition 7: As for Propositions 3 and 4, both behavior and welfare losses depend only on $\bar{x}_t \equiv x_t - y_t$ and not on the specific values of x_t and y_t. For notational simplicity, therefore, this proof shall assume $y_t = 0$ and $x_t = \bar{x}_t$ for all t.

(1) Because committers behave optimally from period 2 onward, committers suffer welfare losses only if they hit in period 1 while TCs never hit. Moreover, if committers hit in period 1, then their period-2 continuation utility is $U_2(1, \alpha^{tc})$. Hence, committers hit in period 1 only if $\bar{x}_1 + \beta\delta U_2(1, \alpha^{tc}) \geq 0$, and their welfare loss from doing so is $-\bar{x}_1 - \delta U_2(1, \alpha^{tc})$. Because a TC in

period 2 with addiction level $k_2 = 1$ could choose to refrain forever after, which yields continuation utility $-\phi\Delta^H$, by revealed preference $U_2(1, \alpha^{tc}) \geq -\phi\Delta^H$. Because $\bar{x}_1 + \beta\delta U_2(1, \alpha^{tc}) \geq 0$ and $U_2(1, \alpha^{tc}) \geq -\phi\Delta^H$ imply $-\bar{x}_1 - \delta U_2(1, \alpha^{tc}) \leq \delta(1-\beta)\phi\Delta^H$, we can conclude $\max_{(\bar{x}_1, \bar{x}_2, ...) \in \mathbf{R}^\infty} \left[WL(0, \alpha^{\text{commit}})\right] \leq \delta(1-\beta)\phi\Delta^H$.

It remains to prove we can hit this bound for any ϕ. To do so, simply let $\bar{x}_1 = \beta\delta\phi\Delta^H$ and let $\bar{x}_t$ be sufficiently small for all $t \geq 2$ that $\alpha^{tc}(k, t) = 0$ for all k and $t \geq 2$ (recall $\bar{x}_t$ can be negative). $\bar{x}_1 = \beta\delta\phi\Delta^H < \delta\phi\Delta^H$ implies committers hit once and TCs never hit, and committers therefore suffer a welfare loss of $-\bar{x}_1 + \delta\phi\Delta^H = \delta(1-\beta)\phi\Delta^H$.

(2) Choose $(\bar{x}_1, \bar{x}_2, ...)$ such that $\bar{x}_1 = \beta\delta\phi\Delta^H$, $\bar{x}_2 + \sigma = \beta\delta\phi\Delta^H$, $\bar{x}_3 + \sigma(1+\gamma) = \beta\delta\phi\Delta^H$, and so forth. $\bar{x}_1 = \beta\delta\phi\Delta^H$ implies a naif in period 1 with $k_1 = 0$ just prefers hitting once to never hitting, and $\bar{x}_2 + \sigma = \beta\delta\phi\Delta^H$ implies a naif in period 2 with $k_2 = 1$ just prefers hitting once to never hitting, and $\bar{x}_3 + \sigma(1+\gamma) = \beta\delta\phi\Delta^H$ implies a naif in period 3 with $k_3 = 1 + \gamma$ just prefers hitting once to never hitting, and so forth. Hence, with this $(\bar{x}_1, \bar{x}_2, ...)$ naifs hit always while TCs never hit, and so naifs suffer a welfare loss of $-\sum_{t=1}^{\infty} \delta^{t-1}\bar{x}_t + \frac{\delta\Delta^H}{1-\delta}$. It is straightforward to derive $\sum_{t=1}^{\infty} \delta^{t-1}\bar{x}_t = \frac{\beta\delta\phi\Delta^H}{1-\delta} - \frac{\delta(\phi-1)\Delta^H}{1-\delta}$, and hence naifs suffer welfare loss $-\sum_{t=1}^{\infty} \delta^{t-1}\bar{x}_t + \frac{\delta\Delta^H}{1-\delta} = \frac{\delta(1-\beta)\phi\Delta^H}{1-\delta}$.

The $(\bar{x}_1, \bar{x}_2, ...)$ chosen above minimize $\sum_{t=1}^{\infty} \delta^{t-1}\bar{x}_t$ subject to naifs planning every period to hit once. The welfare losses cannot be larger because if in some period τ naifs plan to hit $m > 1$ times, then the $\bar{x}'_t$ for periods $\tau + 1$ through $\tau + m$ must be sufficiently large that TCs would hit, which would clearly mean smaller welfare losses.

QED

Proof of Proposition 8: (1) If a TC never hits despite the traumatic event, his utility would be $\sum_{t=1}^{\infty} \delta^{t-1}(-y_t) = -\left(\sum_{t=1}^{N} \delta^{t-1} y_t\right)$. By revealed preference, if he hits some during the traumatic event and perhaps beyond, doing so must yield larger utility, and therefore we can conclude that $\min_{(\rho,\sigma) \in \mathbf{R}^2_+} \left[U_1(0, \alpha^{TC})\right] = -\left(\sum_{t=1}^{N} \delta^{t-1} y_t\right)$.

(2) $V_t(k, \mathbf{r})$ is the long-run continuation utility from refraining forever after. We know that $\alpha^s(0, t) = 0$ for all $t \geq N + 1$. If $\alpha^s(0, N) = 0$, then $U_N(0, \alpha^s) = V_N(0, \mathbf{r})$. If $\alpha^s(0, N) = 1$, then $0 + \beta\delta U_{N+1}(1, \alpha^s) \geq -y_N + \beta\delta V_{N+1}(0, \mathbf{r})$. But since $\alpha^s(0, N) = 1$ implies $U_N(0, \alpha^s) = 0 + \delta U_{N+1}(1, \alpha^s)$, and since $-y_N + \beta\delta V_{N+1}(0, \mathbf{r}) = -(1-\beta)y_N + \beta V_N(0, \mathbf{r})$, it follows that if $\alpha^s(0, N) = 1$ then $U_N(0, \alpha^s) \geq V_N(0, \mathbf{r}) - \frac{1-\beta}{\beta} y_N$. Hence, whether $\alpha^s(0, N) = 0$ or $\alpha^s(0, N) = 1$,

we have $U_N(0,\alpha^s) \geq V_N(0,\mathbf{r}) - \frac{1-\beta}{\beta} y_N$.

Consider period $N-1$. If $\alpha^s(0,N-1) = 0$, then $U_{N-1}(0,\alpha^s) = -y_{N-1} + \delta U_N(0,\alpha^s) \geq V_{N-1}(0,\mathbf{r}) - \frac{1-\beta}{\beta}\delta y_N$. If $\alpha^s(0,N-1) = 1$, then $0 + \beta\delta U_N(1,\alpha^s) = \beta U_{N-1}(0,\alpha^s) \geq -y_{N-1} + \beta\delta U_N(0,\alpha^s) \geq -y_{N-1} + \beta\delta\left[V_N(0,\mathbf{r}) - \frac{1-\beta}{\beta} y_N\right]$, which yields $U_{N-1}(0,\alpha^s) \geq V_{N-1}(0,\mathbf{r}) - \frac{1-\beta}{\beta}(y_{N-1} + \delta y_N)$. Hence, whether $\alpha^s(0,N-1) = 0$ or $\alpha^s(0,N-1) = 1$, we have $U_{N-1}(0,\alpha^s) \geq V_{N-1}(0,\mathbf{r}) - \frac{1-\beta}{\beta}(y_{N-1} + \delta y_N)$.

Iterating this logic, and the fact that $V_1(0,\mathbf{r}) = -\left(\sum_{t=1}^N \delta^{t-1} y_t\right)$, it follows that $U_1(0,\alpha^s) \geq V_1(0,\mathbf{r}) - \frac{1-\beta}{\beta}\left(\sum_{t=1}^N \delta^{t-1} y_t\right) = -\left(\frac{1}{\beta}\sum_{t=1}^N \delta^{t-1} y_t\right)$.

QED

Proof of Proposition 9: (1) For any $\mathbf{p} \equiv (p_1, p_2, ...)$, define $k_1^*(\mathbf{p})$ to be the period-1 addiction level such that a TC is indifferent between hitting always and never hitting. $k_1^*(\mathbf{p})$ is defined by

$$\sum_{t=1}^{\infty} \delta^{t-1}\left[Y_t - p_t + f\left(\gamma^{t-1}k_1^*(\mathbf{p}) + \sum_{n=1}^{t-1}\gamma^{n-1}\right)\right] = \sum_{t=1}^{\infty}\delta^{t-1}\left[Y_t + g\left(\gamma^{t-1}k_1^*(\mathbf{p})\right)\right],$$

which we can rewrite as $\sum_{t=1}^{\infty}\delta^{t-1}(-p_t) + \Phi(k_1^*(\mathbf{p})) = 0$, where $\Phi(k) \equiv \sum_{t=1}^{\infty}\delta^{t-1}\left[f\left(\gamma^{t-1}k + \sum_{n=1}^{t-1}\gamma^{n-1}\right) - g\left(\gamma^{t-1}k\right)\right]$. It is straightforward to show that $\Phi'(k) < 0$.

Define $\bar{\mathbf{p}} \equiv (\bar{p}, \bar{p}, ...)$. Applying Lemma 5, $\bar{k}_t^{tc} = k_1^*(\bar{\mathbf{p}})$ for all t. A simple application of the implicit function theorem yields $d\bar{k}_1^{tc}/d\bar{p} = (1/(1-\delta))/[-\Phi'(k_1^*(\bar{\mathbf{p}}))]$.

Consider next an immediate temporary price change. Given $p_t = \bar{p}$ for all $t \geq 2$, $\bar{k}_t^{tc} = k_1^*(\bar{\mathbf{p}})$ for all $t \geq 2$, which implies that for any $k_1 \in [(k_1^*(\bar{\mathbf{p}}) - 1)/\gamma, k_1^*(\bar{\mathbf{p}})/\gamma]$ the person compares hitting always to never hitting. Hence, for p_1 sufficiently close to $\bar{p}$ that $\bar{k}_1^{tc} \in [(k_1^*(\bar{\mathbf{p}}) - 1)/\gamma, k_1^*(\bar{\mathbf{p}})/\gamma]$, $\bar{k}_1^{tc}$ is determined by the condition $\sum_{t=1}^{\infty}\delta^{t-1}(-p_t) + \Phi(k_1^*(\mathbf{p})) = 0$. At $p_1 = \bar{p}$, $d\bar{k}_1^{tc}/dp_1 = 1/[-\Phi'(k_1^*(\bar{\mathbf{p}}))]$.

Consider a temporary price change in period τ. The logic above implies that at $p_\tau = \bar{p}$, $d\bar{k}_\tau^{tc}/dp_\tau = 1/[-\Phi'(k_1^*(\bar{\mathbf{p}}))]$. Moreover, for p_τ sufficiently close to $\bar{p}$ that $\bar{k}_\tau^{tc} \in [\gamma k_1^*(\bar{\mathbf{p}}), \gamma k_1^*(\bar{\mathbf{p}}) + 1]$, $\bar{k}_{\tau-1}^{tc}$ is determined by the condition $\sum_{n=0}^{\infty}\delta^n(-p_{\tau-1+n}) + \Phi(k_1^*(\mathbf{p})) = 0$ and therefore at $p_\tau = \bar{p}$, $d\bar{k}_{\tau-1}^{tc}/dp_\tau = \delta/[-\Phi'(k_1^*(\bar{\mathbf{p}}))]$. Iterating this logic, we conclude that $\bar{k}_1^{tc}$ is determined by the condition $\sum_{t=1}^{\infty}\delta^{t-1}(-p_t) + \Phi(k_1^*(\mathbf{p})) = 0$, and therefore at $p_\tau = \bar{p}$, $d\bar{k}_1^{tc}/dp_\tau = \delta^{\tau-1}/[-\Phi'(k_1^*(\bar{\mathbf{p}}))]$.

Finally, it follows immediately from above that $(d\bar{k}_1^{tc}/dp_\tau)/(d\bar{k}_1^{tc}/dp_1) = \delta^{\tau-1}$ and $(d\bar{k}_1^{tc}/d\bar{p})/(d\bar{k}_1^{tc}/dp_1) = 1/(1-\delta)$.

(2a) For any $\mathbf{p} \equiv (p_1, p_2, ...)$, define $k_1^{\beta}(\mathbf{p})$ to be the period-1 addiction level such that a naif is

indifferent between hitting always and never hitting. $k_1^\beta(\mathbf{p})$ is defined by

$$\begin{aligned}
&\left[Y_1 - p_1 + f(k_1^\beta(\mathbf{p}))\right] + \beta \sum_{t=2}^{\infty} \delta^{t-1} \left[Y_t - p_t + f\left(\gamma^{t-1} k_1^\beta(\mathbf{p}) + \sum_{n=1}^{t-1} \gamma^{n-1}\right)\right] \\
= \quad &\left[Y_1 + g(k_1^\beta(\mathbf{p}))\right] + \beta \sum_{t=2}^{\infty} \delta^{t-1} \left[Y_t + g\left(\gamma^{t-1} k_1^\beta(\mathbf{p})\right)\right],
\end{aligned}$$

which we can rewrite as $p_1 + \beta \sum_{t=2}^{\infty} \delta^{t-1}(-p_t) + \tilde{\Phi}(k_1^\beta(\mathbf{p})) = 0$, where $\tilde{\Phi}(k) \equiv f(k) - g(k) + \beta \sum_{t=2}^{\infty} \delta^{t-1} \left[f\left(\gamma^{t-1}k + \sum_{n=1}^{t-1} \gamma^{n-1}\right) - g\left(\gamma^{t-1}k\right)\right]$. It is straightforward to show that $\tilde{\Phi}'(k) < 0$.

Applying Lemma 5, if $k^*(\beta, \bar{p}) < \tilde{k}(\beta, \bar{p})$ then $\bar{k}_t^n = k_1^\beta(\bar{\mathbf{p}})$ for all t. It is straightforward to show that for small price changes $\bar{k}_1^n$ is still the addiction level at which the person is indifferent between hitting always and never hitting (the logic is the same as that used in the proof of part 1). Applying the implicit function theorem yields $d\bar{k}_1^n/d\bar{p} = (1 + \beta\delta/(1-\delta))/[-\tilde{\Phi}'\left(k_1^\beta(\bar{\mathbf{p}})\right)]$, $d\bar{k}_1^n/dp_1 = 1/[-\tilde{\Phi}'\left(k_1^\beta(\bar{\mathbf{p}})\right)]$, and $d\bar{k}_1^n/dp_\tau = \left(\beta\delta^{\tau-1}\right)/[-\tilde{\Phi}'\left(k_1^\beta(\bar{\mathbf{p}})\right)]$. The result follows.

(2b) For any $\mathbf{p} \equiv (p_1, p_2, ...)$, define $\tilde{k}_1^\beta(\mathbf{p})$ to be the period-1 addiction level such that a naif is indifferent between hitting once and never hitting. $\tilde{k}_1^\beta(\mathbf{p})$ is defined by

$$\begin{aligned}
&\left[Y_1 - p_1 + f(\tilde{k}_1^\beta(\mathbf{p}))\right] + \beta \sum_{t=2}^{\infty} \delta^{t-1} \left[Y_t + g\left(\gamma^{t-1} \tilde{k}_1^\beta(\mathbf{p}) + \gamma^{t-2}\right)\right] \\
= \quad &\left[Y_1 + g(\tilde{k}_1^\beta(\mathbf{p}))\right] + \beta \sum_{t=2}^{\infty} \delta^{t-1} g\left(\gamma^{t-1} \tilde{k}_1^\beta(\mathbf{p})\right),
\end{aligned}$$

which we can rewrite as $p_1 + \tilde{\Gamma}(k_1^\beta(\mathbf{p})) = 0$, where $\tilde{\Gamma}(k) \equiv f(k) - g(k) +$
$\beta \sum_{t=2}^{\infty} \delta^{t-1} \left[g\left(\gamma^{t-1}k + \gamma^{t-2}\right) - g\left(\gamma^{t-1}k\right)\right]$. It is straightforward to show that $\tilde{\Gamma}'(k) < 0$.

Applying Lemma 5, if $\tilde{k}(\beta, \bar{p}) < k^*(\beta, \bar{p})$ then $\bar{k}_t^n = \tilde{k}_1^\beta(\bar{\mathbf{p}})$ for all t. It is again straightforward to show that for small price changes $\bar{k}_1^n$ is still the addiction level at which the person is indifferent between hitting once and never hitting. Applying the implicit function theorem yields $d\bar{k}_1^n/d\bar{p} = 1/[-\tilde{\Gamma}'\left(\tilde{k}_1^\beta(\bar{\mathbf{p}})\right)]$, $d\bar{k}_1^n/dp_1 = 1/[-\tilde{\Gamma}'\left(\tilde{k}_1^\beta(\bar{\mathbf{p}})\right)]$, and $d\bar{k}_1^n/dp_\tau = 0$. The result follows.

QED

References

Ainslie, G. (1975). "Specious reward: A behavioral theory of impulsiveness and impulse control." *Psychological Bulletin*, **82**, 463-496.

Ainslie, G. (1991). "Derivation of 'Rational' Economic Behavior from Hyperbolic Discount Curves." *American Economic Review*, **81**, 334-340.

Ainslie, G. (1992). *Picoeconomics: The strategic interaction of successive motivational states within the person*. New York: Cambridge University Press.

Ainslie, G. and N. Haslam (1992*a*). "Self-control," in G. Loewenstein and J. Elster, eds., *Choice Over Time*. New York: Russell Sage Foundation, 177-209.

Ainslie, G. and N. Haslam (1992*b*). "Hyperbolic Discounting," in G. Loewenstein and J. Elster, eds., *Choice Over Time*. New York: Russell Sage Foundation, 57-92.

Akerlof, G. A. (1991). "Procrastination and Obedience." *American Economic Review*, **81**, 1-19.

Becker, G. S. and K. M. Murphy (1988). "A Theory of Rational Addiction." *Journal of Political Economy*, **96**, 675-700.

Becker, G. S., M. Grossman, and K. M. Murphy (1994). "An Empirical Analysis Of Cigarette Addiction." *American Economic Review*, **84**, 396-418.

Caillaud, B., D. Cohen, and B. Jullien (1996). "Towards a Theory of Self-Restraint." Mimeo, CEPREMAP.

Carrillo, J. (1999). "Self-Control, Moderate Consumption, and Craving." CEPR Discussion Paper 2017.

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

Elster, J., ed. (1999). *Addiction: Entries and Exits*. New York: Russell Sage Foundation.

Fischer, C. (1997). "Read This Paper Even Later: Procrastination with Time-Inconsistent Preferences." Dissertation, University of Michigan.

Goldbaum, D. (2000). "Life Cycle Consumption of a Harmful and Addictive Good." *Economic Inquiry*, **38**, 458-469.

Goldman, S. M. (1979). "Intertemporally Inconsistent Preferences and the Rate of Consumption." *Econometrica*, **47**, 621-626.

Gruber, J. and B. Koszegi (2000). "Is Addiction 'Rational'? Theory and Evidence." NBER Working Paper 7507.

Harris, C. and D. Laibson (forthcoming). "Dynamic Choices of Hyperbolic Consumers." *Econometrica*.

Herrnstein, R.J., G. Loewenstein, D. Prelec, and W. Vaughan, Jr. (1993). "Utility Maximization and Melioration: Internalities in Individual Choice." *Journal of Behavioral Decision Making*, **6**, 149-185.

Laibson, D. (1994). "Essays in Hyperbolic Discounting." Economics, MIT.

Laibson, D. (1996). "Hyperbolic Discount Functions, Undersaving, and Savings Policy." NBER Working Paper 5635.

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

Loewenstein, G., T. O'Donoghue, and M. Rabin (2000). "Projection Bias in Predicting Future Utility." U.C. Berkeley Economics Department Working Paper E00-284.

Loewenstein, G. and D. Prelec (1992). "Anomalies in Intertemporal Choice: Evidence and an Interpretation." *Quarterly Journal of Economics*, **107**, 573-597.

O'Donoghue, T. and M. Rabin (1999*a*). "Doing It Now or Later." *American Economic Review*, **89**, 103-124.

O'Donoghue, T. and M. Rabin (1999*b*). "Addiction and Self Control," in J. Elster, ed., *Addiction: Entries and Exits*. New York: Russell Sage Foundation, 169-206.

O'Donoghue, T. and M. Rabin (1999*c*). "Procrastination in Preparing for Retirement," in Henry Aaron, ed., *Behavioral Dimensions of Retirement Economics*. Washington D.C. and New York: Brookings Institution Press & Russell Sage Foundation, 125-156.

O'Donoghue, T. and M. Rabin (2001), "Choice and Procrastination," *Quarterly Journal of Economics*, **116**(1), 121-160.

Orphanides, A. and D. Zervos (1995). "Rational Addiction with Learning and Regret." *Journal of Political Economy*, **103**, 739-758.

Phelps, E. S. and R. A. Pollak (1968). "On Second-best National Saving and Game-equilibrium Growth." *Review of Economic Studies*, **35**, 185-199.

Pollak, R. A. (1968). "Consistent Planning." *Review of Economic Studies*, **35**, 201-208.

Pollak, R. A. (1970). "Habit Formation and Dynamic Demand Functions." *Journal of Political Economy*, **78**, 745-763.

Ryder, H. E. and G.M. Heal (1973). "Optimal Growth with Intertemporally Dependent Preferences." *Review of Economic Studies*, **40**, 1-33.

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

Suranovic, S., R. Goldfarb, and T. Leonard (1999). "An Economic Theory of Cigarette Addiction." *Journal of Health Economics*, **18**, 1-29.

Thaler, R. H. (1991). "Some Empirical Evidence on Dynamic Inconsistency," in *Quasi Rational Economics*. New York: Russell Sage Foundation, 127-133.

Thaler, R. and G. Loewenstein (1992). "Intertemporal Choice," in R. Thaler, ed., *The Winner's Curse: Paradoxes and Anomalies of Economic Life*. New York: Free Press, 92-106.

Wang, R. (1997). "The Optimal Consumption and the Quitting of Harmful Addictive Goods." Mimeo, Queens University.