2-1-2017

Behavioral Characterizations of Naiveté for Time-Inconsistent Preferences

David S. Ahn

Ryota Iijima

Yves Le Yaouanq

Todd Sarver

BEHAVIORAL CHARACTERIZATIONS OF NAIVETÉ FOR TIME-INCONSISTENT PREFERENCES

By

David S. Ahn, Ryota Iijima, Yves Le Yaouanq, and Todd Sarver

February 2017

COWLES FOUNDATION DISCUSSION PAPER NO. 2074

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Behavioral Characterizations of Naiveté for Time-Inconsistent Preferences*

David S. Ahn† Ryota Iijima‡ Yves Le Yaouanq§ Todd Sarver¶

First Version: April 28, 2015

Current Draft: November 14, 2016

Abstract

We propose nonparametric definitions of absolute and comparative naiveté. These definitions leverage ex-ante choice of menu to identify predictions of future behavior and ex-post (random) choices from menus to identify actual behavior. The main advantage of our definitions is their independence from any assumed functional form for the utility function representing behavior. An individual is sophisticated if she is indifferent between choosing from a menu ex post or committing to the actual distribution of choices from that menu ex ante. She is naive if she prefers the flexibility in the menu, reflecting a mistaken belief that she will act more virtuously than she actually will. We propose two definitions of comparative naiveté and explore the restrictions implied by our definitions for several prominent models of time inconsistency. Finally, we discuss the implications of general naiveté for welfare and the design of commitment devices.

KEYWORDS: Naive, sophisticated, time inconsistent, comparative statics

*Ahn and Sarver acknowledge the financial support of the National Science Foundation through Grants SES-1357719 and SES-1357955. We also thank Ned Augenblick, Roland Bénabou, Drew Fudenberg, Christian Gollier, David Laibson, Bart Lipman, Takeshi Murooka, Jawwad Noor, Wolfgang Pesendorfer, Matthew Rabin, Philipp Sadowski, Ran Spiegler, Tomasz Strzalecki, Norio Takeoka, Jean Tirole, and seminar participants at Arizona State, Berkeley, Bocconi, Caltech, Columbia, Duke, Harvard, ITAM, Kansas, London School of Economics, Michigan, Princeton, Queen Mary, UC Riverside, Rochester, and Washington University in St. Louis for helpful comments and discussions. This paper incorporates results previously circulated under the titles "Comparative Measures of Naiveté" by Ahn, Iijima, and Sarver and "Anticipating Preference Reversals" by Le Yaouanq.

†Department of Economics, University of California, Berkeley, 530 Evans Hall #3880, Berkeley, CA 94720-3880. Email: dahn@econ.berkeley.edu.

‡Cowles Foundation for Research in Economics, Yale University, 30 Hillhouse Ave, New Haven, CT 06510. Email: ryota.iijima@yale.edu.

§Seminar für Organisationsökonomik, Ludwig-Maximilians-Universität, Kaulbachstr. 45, München 81825, Germany. Email: yves.leyaouanq@econ.lmu.de.

¶Department of Economics, Duke University, 213 Social Sciences/Box 90097, Durham, NC 27708. Email: todd.sarver@duke.edu.

1 Introduction

Models of dynamic inconsistency play an important role in a wide-ranging set of economic applications, and there is strong and increasing interest in the implications of naiveté when individuals mispredict their future behavior.[1] While naiveté often yields surprising and significant consequences, so far these effects are usually understood within the context of specific utility representations, where the existence and comparison of naiveté are defined and tested through parameters like discount factors or probabilities.

In this paper, we introduce general nonparametric definitions of naiveté and sophistication, as well as comparative measures of naiveté. We then characterize the implications of these definitions for a broad class of utility specifications. Our behavioral definitions leverage two pieces of choice data. First, we use preference for commitment to measure *anticipated* behavior from an ex-ante perspective before the realization of temptation. Formally, the individual's preferences over different option sets (or menus) capture her demand for commitment and allow an inference of her beliefs regarding her future behavior. Second, we use choices from option sets to measure *actual* behavior from an ex-post perspective under the influence of temptation and after the level of commitment is fixed. Since uncertainty about future behavior seems especially compelling under naiveté and is increasingly relevant in applied work, we formally accommodate this uncertainty by modeling ex-post behavior as a random choice rule.

For a simple illustration of our approach, consider first an individual who makes deterministic choices. Her ex-ante ranking of option sets is given by a preference $\succsim$, and her ex-post choice from any menu is given by a choice function $\mathcal{C}$.[2] When choosing between two options p and q, an individual may prefer p if committing ex ante, $\{p\} \succ \{q\}$, yet choose q if given the option ex post, $\mathcal{C}(\{p, q\}) = q$. This pattern is indicative of time inconsistency and has been documented in numerous contexts, e.g., a preference to maintain a healthy diet, decrease spending, or engage timely effort in a difficult task that goes unfulfilled ex post. Still, additional information is needed to determine whether the individual is sophisticated or naive about this inconsistency. If we also observe a strict preference to retain the option p ex ante, $\{p, q\} \succ \{q\}$, then we can further infer that she (incorrectly) anticipates that p will be her ex-post choice from the menu $\{p, q\}$ and hence she is naive. In the more general case of stochastic choice, if p is chosen with probability

[1] A recent survey of empirical applications can be found in Section 2.1 of DellaVigna (2009) and a survey of some theoretical applications in contract theory can be found in Koszegi (2014).

[2] We focus throughout the paper on choice functions rather than correspondences, which presumes the individual uses some tie-breaking procedure to select between equally attractive options. Our primitives for stochastic choice make similar implicit assumptions. Importantly, our results do not depend in any way on how ties are broken. Hence, while our results can easily be extended to deal with choice correspondences (and their stochastic generalizations), it is a strength of the current analysis that knowledge of the complete set of possible options that the individual is willing to choose from a menu is not required.

α from the menu $\{p, q\}$ at the ex-post stage, then the relevant ex-ante comparison is between the menu $\{p, q\}$ and commitment to the mixture $\{\alpha p + (1 - \alpha)q\}$. A strict preference for the former indicates biased beliefs that overestimate the probability of choosing the ex-ante more appealing alternative p.

Our behavioral definitions extend the same approach to arbitrary choice sets. To test absolute naiveté and sophistication, we compare an individual's predicted value for a menu x of different options against the actual value of her ex-post choice $\mathcal{C}(x)$ from that menu. Ex ante, a sophisticate correctly anticipates her future choice and is indifferent between maintaining the flexibility to choose from x later or committing to her eventual choice $\mathcal{C}(x)$ now, i.e., $x \sim \{\mathcal{C}(x)\}$. In contrast, a naif mistakenly anticipates making a more virtuous choice and prefers to maintain the flexibility in x, i.e., $x \succsim \{\mathcal{C}(x)\}$. In the case of uncertain temptations and random choice, we maintain this basic intuition by comparing her preference for the menu versus committing to the lottery over outcomes induced by her distribution of choices. As we discuss later in the introduction, our definitions are closely related to several recent empirical studies of time inconsistency and naiveté.

While the behavioral implications of absolute naiveté have received some attention in the literature, the behavior associated with increases in naiveté has not been nearly so well explored—especially in the case of stochastic choice. As a result, even within specific models, the proper parametric restrictions that capture increased naiveté are not fully understood or agreed upon. To shed some light on this issue, we propose two behavioral definitions of comparative naiveté. For ease of illustration, consider first the special case of deterministic choice. Our first definition compares beneficial commitment opportunities that are naively declined. A commitment to the singleton menu $\{p\}$ is beneficial if $\{p\} \succsim \{\mathcal{C}(x)\}$, that is, if p is more virtuous than the outcome $\mathcal{C}(x)$ that would be chosen from x. A naive agent may nonetheless prefer x to $\{p\}$; that is, instead of taking the opportunity to commit to $\{p\}$, she maintains the flexibility of x, anticipating making a more virtuous choice, but ends up with the more indulgent $\mathcal{C}(x)$. So a beneficial commitment is declined if $x \succsim \{p\} \succsim \{\mathcal{C}(x)\}$. Our first definition is that an individual is more naive than another if she declines more advantageous commitments.

Our second definition compares individuals' anticipated and actual indirect utilities for a menu. A naive individual overvalues flexibility. Correspondingly, our second proposal is that an individual is more naive than another if the difference between her believed and actual indirect utilities for a menu is always larger. We provide a primitive behavioral condition that characterizes this comparison. We prove that this notion is less demanding than our first comparative measure and hence more completely ranks naiveté across individuals. In the case of random choice, both comparative definitions extend by replacing the deterministic choice with the induced lottery over outcomes.

Using one of the most comprehensive models of time-inconsistent preferences available, the *random Strotz representation*, we show that our definitions of absolute and comparative naiveté characterize sharp and intuitive parametric restrictions. As we will illustrate using examples and applications throughout the paper, this representation is general enough to include the majority of all utility representations for time-inconsistent preferences that appear in the applied literature.[3] Our approach therefore unifies disparate models in the literature and illuminates a basic common behavior that undergirds their evaluations of naiveté: underdemand for commitment.

As an illustration, consider the following stochastic generalization of the Strotzian quasi-hyperbolic representation. An individual would like to choose a consumption stream to maximize her exponentially discounted stream of instantaneous utility u with discount factor δ. Instead, future utility is discounted against the present by an additional present-bias factor β that is random and follows a distribution F on $[0,1]$. Her possibly mistaken belief is that her present-bias will instead follow distribution $\hat{F}$. For this model, our absolute definition of naiveté turns out to be equivalent to the first-order stochastic dominance relation $\hat{F} \geq_{FOSD} F$, i.e., $\hat{F}(\beta) \leq F(\beta)$ for all $\beta \in [0,1]$. A naive individual is therefore overoptimistic about her virtue in the statistical sense of overweighting more patient present-bias factors. In addition, individual 1 is more naive than individual 2 in our first stronger sense if and only if $\hat{F}_1 \geq_{FOSD} \hat{F}_2 \geq_{FOSD} F_2 \geq_{FOSD} F_1$. That is, a more naive agent has more optimistic beliefs ($\hat{F}_1 \geq_{FOSD} \hat{F}_2$) while simultaneously engaging in less virtuous behavior ($F_2 \geq_{FOSD} F_1$). Under our weaker second comparison, individual 1 is more naive than individual 2 if and only if $F_1(\beta) - \hat{F}_1(\beta) \geq F_2(\beta) - \hat{F}_2(\beta)$ for all $\beta \in [0,1]$. In other words, the more naive individual underestimates the probability of greater impatience (low values of β) by more than the less naive individual.

This general stochastic representation has two important special cases. First, suppose $\hat{F}$ and F are supported on β and 1. That is, there is some chance an individual succumbs to present-bias β and some chance she takes the virtuous action and maximizes exponential discounted utility. In actuality, the chance of being virtuous is $\theta = 1 - F(\beta)$, but the decision maker thinks the chance of being virtuous is $\hat{\theta} = 1 - \hat{F}(\beta)$. This corresponds to a model of frequency naiveté originally proposed by Eliaz and Spiegler (2006). Absolute naiveté in this special case is equivalent to $\hat{\theta} \geq \theta$. Individual 1 is more naive than 2 in our stronger first sense if and only if $\beta_1 = \beta_2$ and $\hat{\theta}_1 \geq \hat{\theta}_2 \geq \theta_2 \geq \theta_1$. Thus both individuals share the same potential levels of realized present-bias, but the more naive one believes she is less likely to succumb to temptation ($\hat{\theta}_1 \geq \hat{\theta}_2$) while in reality she is more likely to be present-biased ($\theta_2 \geq \theta_1$). Our weaker second comparison requires either that 2 is sophisticated (so $\beta_2 = 1$ or $\hat{\theta}_2 = \theta_2$) or that $\beta_2 \geq \beta_1$ and $\hat{\theta}_1 - \theta_1 \geq \hat{\theta}_2 - \theta_2$, which is

[3] One important exception is models that incorporate costly self-control. We apply our definitions to the random self-control representation as an extension in Section 7.2, and we explore alternative definitions of naiveté for self-control preferences in a companion paper Ahn, Iijima, and Sarver (2016).

more general since the level of present-bias can be strictly more severe for the more naive individual and only the differences in their beliefs need to be ordered.

As a second special case, suppose $\hat{F}$ and F are deterministic and concentrated respectively on $\hat{\beta}$ and β. This is the naive quasi-hyperbolic model introduced by O'Donoghue and Rabin (2001). Then an individual is naive if and only if $\hat{\beta} \geq \beta$. An individual is strongly more naive then another if and only if $\hat{\beta}_1 \geq \hat{\beta}_2 \geq \beta_2 \geq \beta_1$. However, in contrast to the prior special case with more optimistic probability weights, the weaker second definition of comparative naiveté is here equivalent to the first, excepting the case where individual 2 is sophisticated ($\hat{\beta}_2 = \beta_2$). In particular, $\hat{\beta}_1 - \beta_1 \geq \hat{\beta}_2 - \beta_2$ does not imply individual 1 is weakly more naive.

As these cases show, our approach can bring to light unifying themes and subtle distinctions across models. But beyond improved theoretical understanding, behavioral definitions of naiveté provide relevant substantive benefits. They permit an examination of which positive predictions in applications rely on functional-form assumptions and which predictions are inherent features of naiveté. For example, a more risk-accepting investor will always choose a risky equity position over a risk-free bond whenever a more risk-averse investor does. Similarly, we can ask whether predictions regarding savings or procrastination are artifacts of an assumed utility or are robust implications of naiveté. In turn, a deeper understanding of the mechanics of naive choice also improves normative analysis. In particular, effective design of commitment devices can hinge crucially on the assumed level of sophistication. Duflo, Kremer, and Robinson (2011) examine a theoretical model where the optimal timing of when to offer a commitment depends on whether individuals are sophisticated or naive regarding the degree of their present bias, and they provide evidence from Kenyan fertilizer adoption that individuals are naive and would benefit from earlier and time-limited commitments. Nonparametric definitions of naiveté provide a language broad enough to understand the consequence of policy interventions when citizens have qualitatively different forms of naiveté and are best approximated by a variety of formal models, and to understand which policies work for which assumed models.

Our use of ex-ante and ex-post behavior has several precedents in recent empirical studies of time inconsistency and naiveté. For example, DellaVigna and Malmendier (2006) study both the choice of gym membership, which determines the feasible set of attendance/payment pairs, and subsequent attendance levels; Shui and Ausubel (2005) observe consumers' choices of credit card contracts and their subsequent borrowing behavior; Giné, Karlan, and Zinman (2010) offer subjects commitment contracts that incentivize smoking cessation and later test whether or not the subjects smoked; Kaur, Kremer, and Mullainathan (2015) allow subjects to choose wage contracts that constrain their feasible future effort/consumption pairs and then observe actual effort ex post; Augenblick, Niederle, and Sprenger (2015) ask subjects to choose an intertemporal allocation

of effort and a probability of being committed to it and then observe whether subjects wish to revise that plan when the first date of task completion arrives. Not only do these papers use similar choice data, but those that test for naiveté identify it using behavior that is closely related to our definition. In fact, if individuals satisfy a basic dominance condition—they prefer more money to less—then the evidence of naiveté found in several of these papers can be mapped exactly into our definition. For example, purchasing an unlimited gym membership and failing to attend the gym would be classified as naive under our definition, since purchasing the membership is revealed ex-ante preferred to not joining the gym, which is preferred by dominance to committing to pay for a membership and not attend.[4]

There are also papers in decision theory that use behavior at different time periods to capture sophistication under time inconsistency, as surveyed by Lipman and Pesendorfer (2013). Noor (2011) considers preferences over a recursive domain that includes ex-ante and ex-post choice preferences as projections; he pioneered the approach of using temporal choice as a domain for explicitly testing the sophistication implicitly assumed in most ex-ante axiomatic models of temptation. Kopylov (2012) relaxes Noor's sophistication condition and considers agents who choose flexibility ex ante that is subsequently unused ex post. Kopylov eschews mistaken or naive beliefs, but rather interprets the relaxation of sophistication as reflecting a direct psychic benefit of maintaining positive self-image. Finally, Dekel and Lipman (2012) observe that ex-ante and ex-post choice can be combined to empirically distinguish random Strotz representations from others that involve costly self-control. Much of the technical apparatus from Dekel and Lipman (2012) ends up being useful in studying naiveté, as we will explain in the body of the paper.

The next section describes our formal primitives. Section 3 introduces our absolute definition of naiveté. We begin with the special deterministic case to introduce and ground concepts, and then move on to the general random case. Section 4 introduces our strong and weak comparisons of naiveté. In both sections, we explore the implications of these absolute and comparative definitions for general random Strotz representations. In Section 5, we examine several popular specifications of dynamic inconsistency, such as quasi-hyperbolic discounting and general diminishing impatience, and establish the parametric restrictions implied by our definitions in these special cases. Section 6 applies our setup to analyze the general welfare implications of naiveté for policies that introduce

[4]Two other types of data are also sometimes used as evidence of naiveté: The first is procrastination in completing tasks that have immediate costs and delayed rewards. We discuss in Section 3.1 how procrastination is a special case of our definition of naiveté. The second is surveys that directly ask subjects to predict their future behavior. For example, a recent experiment by Augenblick and Rabin (2015) incentivized direct reports of subjects' predictions of future behavior. Importantly, since prediction-accuracy bonuses allow subjects to use their predictions as soft commitment devices for their future behavior, Augenblick and Rabin (2015) invoke a structural model that allows them to correct for the resulting bias in belief estimates. It is not obvious how to adapt a nonparametric approach like the one in this paper to their data.

new commitment devices. Finally, Section 7 discusses areas where our model could be generalized, including extensions to models of costly self-control and uncertain normative preferences.

2 Primitives

We study a two-stage model with an agent who initially decides a menu of several options and subsequently selects a particular option from that menu.

Let C be a compact and metrizable space of outcomes. Let $\Delta(C)$ denote the set of lotteries (countably-additive Borel probability measures) over C, with typical elements $p, q, \ldots \in \Delta(C)$. When it causes no confusion, we slightly abuse notation and write c in place of the degenerate lottery $\delta_c \in \Delta(C)$ supported on c. Let $\mathcal{K}(\Delta(C))$ denote the family of nonempty compact subsets of $\Delta(C)$ with typical elements $x, y, \ldots \in \mathcal{K}(\Delta(C))$. An *expected-utility function* is a continuous function $u : \Delta(C) \to \mathbb{R}$ such that $u(\alpha p + (1-\alpha)q) = \alpha u(p) + (1-\alpha)u(q)$ for all lotteries p, q. A function is *nontrivial* if it is not constant. We write $u \approx v$ when u and v are expected-utility functions and u is a positive affine transformation of v. For a fixed expected-utility function u and menu x, let $B_u(x) \equiv \operatorname{argmax}_{p \in x} u(p)$.

We consider a pair of behavioral primitives. The first primitive is a preference relation $\succsim$ on $\mathcal{K}(\Delta(C))$, with indifference $\sim$ and strict preference $\succ$ defined as usual. The behavior encoded in $\succsim$ is taken before the direct experience of temptation but while (possibly incorrectly) anticipating its future occurrence. The second primitive is a random choice rule $\lambda : \mathcal{K}(\Delta(C)) \to \Delta(\Delta((C))$ such that $\lambda^x(x) = 1$, where $\Delta(\Delta(C))$ denotes the space of lotteries over $\Delta(C)$. The behavior encoded in λ is taken while experiencing temptation. For each $x \in \mathcal{K}(\Delta(C))$, λ^x is a probability measure over lotteries, with $\lambda^x(y)$ denoting the probability of choosing a lottery in the set $y \subset x$ when the choice set is the menu x. We refer to the first stage of choice of a menu as occurring "ex ante" and the second stage of choice from a menu as occurring "ex post," that is, before and after the realization of temptation.

We sometimes specialize to choice functions without randomization for their substantive importance and expositional clarity. A random choice function λ is *deterministic* if λ^x is degenerate for all menus x, that is, $\lambda^x = \delta_p$ for some $p \in x$. Identifying the Dirac measure δ_p with p itself, we can notate λ as a standard choice function $\mathcal{C} : \mathcal{K}(\Delta(C)) \to \Delta(C)$.[5] In that case, $\mathcal{C}(x) = p$ for $\delta_p = \lambda^x$.

These primitives echo prior work by Ahn and Sarver (2013) on unforeseen contingen-

[5]Recall the final outcomes are themselves lotteries. The determinacy here is in the sense that the decision maker does not randomize her selection among these lotteries.

cies. That paper inferred unawareness of future taste contingencies by comparing choices before and after the realization of subjective uncertainty: Observing ex-ante demand for flexibility and ex-post exercise of flexibility can reveal unawareness and provide positive foundations for the measurement of an unforeseen contingency, while the standard approach of using only ex-ante preferences cannot. Similarly, here we use demand for commitment in the first stage and then indulgence of temptation in the second stage to infer naiveté. Very broadly speaking, under-demand for flexibility can reveal unawareness of future taste contingencies, while under-demand for commitment can reveal naiveté about future temptations.

3 Absolute Naiveté

3.1 Benchmark Case: Deterministic Choice

To facilitate intuition, we begin by specializing attention to the important case of choice without randomization and tabling the general random case until the next subsection. For now, assume a deterministic choice function $\mathcal{C}$. We propose the following definitions of absolute sophistication and naiveté for deterministic choice.

Definition 1. *An individual is* sophisticated *if* $x \sim \{\mathcal{C}(x)\}$ *for all menus* x*. An individual is* naive *if* $x \succsim \{\mathcal{C}(x)\}$ *for all menus* x*. An individual is* strictly naive *if she is naive and not sophisticated.*

A sophisticated individual correctly anticipates choosing $\mathcal{C}(x)$ from x. A naive individual erroneously values the option to make more virtuous choices, thinking her final choice will be more virtuous than $\mathcal{C}(x)$. Many decisions that open or restrict future options can be modeled as menus and can therefore be related to our definitions. For example, as we discussed in the introduction, purchasing an unlimited gym membership can be modeled as the option set that includes any number of monthly visits, each paired with the fixed cost of the membership. Similarly, many financial decisions, like opening a line of credit or putting money into a restricted retirement account, can be viewed as adding or removing options from future decisions. In these examples, we argue that some consumers may strictly prefer x to $\mathcal{C}(x)$, indicating a lack of sophistication in predicting their future choices.

Another set of problems where naiveté can manifest is decisions about the timing of completion of a task. O'Donoghue and Rabin (1999, 2001) explored the theoretical implications of naiveté in this class of problems and found that it can lead to procrastination in completing tasks that have immediate costs and delayed rewards. Their predictions have since been used to explain empirical evidence of procrastination, ranging from delay

in setting up 401(k) accounts with employer matching contributions (Madrian and Shea (2001)) to delay in canceling unused gym memberships (DellaVigna and Malmendier (2006)). Decisions about the timing of task completion are a special case of our framework of choice between and from option sets. To illustrate, let d_1, d_2, d_3 denote doing it now, tomorrow, or in two periods. The choice of whether or not to complete the task in the first period is a choice between the menus $\{d_1\}$ (committing by doing it now) and $\{d_2, d_3\}$ (having the option of doing it tomorrow or delaying again). Procrastination corresponds to $\{d_2, d_3\} \succ \{d_1\} \succ \{d_3\}$ in the first period and $\mathcal{C}(\{d_2, d_3\}) = d_3$ in the second. The individual prefers delaying the task by exactly one period and mistakenly believes that delaying today will result in completion of the task tomorrow. Note that procrastination implies strict naiveté according to our definition, since $\{d_2, d_3\} \succ \{\mathcal{C}(\{d_2, d_3\})\}$.[6] This mapping from procrastination into our definition of naiveté can be generalized to any number of periods by taking the appropriate three-period snapshot: Select any three periods such that on the subtree consisting of only these periods the individual procrastinates on date 1 due to the mistaken belief that she will complete the task on date 2.[7]

In our definition, inferring sophistication from $x \sim \{\mathcal{C}(x)\}$ assumes consequentialism; that is, the individual is indifferent between committing to her (correctly) anticipated choice $\mathcal{C}(x)$ from x at the ex-ante stage or selecting the menu x with the belief that she will choose $\mathcal{C}(x)$ ex post. Put differently, adding or removing unchosen options has no effect on the evaluation of a menu. In contrast, an individual who exerts costly willpower to avoid choosing tempting options as in Gul and Pesendorfer (2001) does not evaluate a menu only by its choice consequences. In this case, she may strictly prefer to remove these unchosen temptations. In Section 7.2, we show that if individuals can exert costly self-control, our behavioral test of naiveté can lead to false negatives but not false positives: Satisfying our definition of naiveté in the presence of costly self-control implies *a fortiori* that the individual is naive; however, satisfying our definition of sophistication does not guarantee that an individual with Gul and Pesendorfer (2001) preferences is in fact sophisticated.[8]

The opposite violation of the suggested indifference for sophistication, where $\{\mathcal{C}(x)\} \succ$

[6] Similarly, by interpreting d_1 as a beneficial commitment opportunity that is naively declined by a procrastinating individual, greater tendency to procrastinate is a special case of the comparative measure of naiveté that will be introduced in Section 4.1.

[7] In a recent paper developed independently of, but subsequent to, previously circulated drafts of this paper (Ahn and Sarver (2015); Le Yaouanq (2015)), Freeman (2016) adopts similar conditions to study naiveté in this special case of deterministic stopping problems and procrastination.

[8] In a companion paper Ahn, Iijima, and Sarver (2016), we modify the definition of sophistication from Noor (2011) to provide a tight behavioral characterization of naiveté for both deterministic self-control preferences and deterministic Strotz preferences. However, the trade-off is that the definition of naiveté in Ahn, Iijima, and Sarver (2016) cannot be extended to random choice, which is a principal objective of the current paper.

x and individuals underestimate their future virtue, is also possible.[9] Many of our results have analogous statements for this case, as recorded in Section S.1 of the Supplemental Appendix. This direction receives less attention and seems less empirically relevant, so the main text focuses on traditional naiveté.

The ubiquitous Strotz model of dynamic inconsistency offers a general application for these concepts. The sophisticated Strotz model is specified by two preferences. The first is her ex-ante commitment preference over future consumption, as represented by the utility function u. The second is her temptation preference that governs her actual consumption choices at the ex-post stage, as represented by the utility function v. Naiveté requires divergence between believed and actual consumption. Specification of a naive Strotz individual therefore requires a third preference to capture her possibly erroneous beliefs about her future behavior, as represented by the utility function $\hat{v}$.[10]

Definition 2. *A* Strotz representation *of* $(\succsim, \mathcal{C})$ *is a triple* $(u, v, \hat{v})$ *of nontrivial expected-utility functions such that the function* $U : \mathcal{K}(\Delta(C)) \to \mathbb{R}$ *defined by*

$$U(x) = \max_{p \in B_{\hat{v}}(x)} u(p)$$

is a utility representation of $\succsim$ *and*

$$\mathcal{C}(x) \in B_u(B_v(x)).$$

While she anticipates maximizing $\hat{v}$, a naive Strotzian agent's ex-post behavior $\mathcal{C}$ actually maximizes v. Note that both the domain of choice and the representation itself are quite general. For example, C could be a set of infinite-horizon consumption streams, and hence quasi-hyperbolic discounting (β-δ preferences) is a special case of the Strotz representation (see Section 5).

The following result demonstrates that the basic definition of naiveté characterizes sharp parametric restrictions on $\hat{v}$ and v. A naive individual believes that her future behavior will be more virtuous than it actually is. For the parameters of the Strotz model, this means that the anticipated utility $\hat{v}$ is more aligned with the commitment utility u than the actual utility v that will govern future consumption. The alignment has a specific structure: $\hat{v}$ is a linear combination of u and v, that is, $\hat{v} \approx \alpha u + (1 - \alpha)v$. The belief $\hat{v}$ puts additional unjustified weight on the normative utility u, but aggregates u with v in a linear manner. This excludes the case where the believed temptation is orthogonal to the actual temptation. For example, our definition excludes an individual

[9] Ali (2011) shows that such a pessimistic belief can arise and persist in a model of Bayesian experimentation.

[10] Recall that a utility function is nontrivial if it is not constant, and $B_v(x)$ was defined as $\operatorname{argmax}_{q \in x} v(q)$.

who actually will be tempted to indulge in sweet treats but believes she will be tempted to indulge in salty treats. This structure also relies crucially on the linear structure of the domain of lotteries and the assumed expected-utility functions.

Definition 3. *Let $u, v, \hat{v}$ be expected-utility functions. Then $\hat{v}$ is* more u-aligned *than v, written as $\hat{v} \gg_u v$, if either $\hat{v} \approx \alpha u + (1-\alpha)v$ for some $\alpha \in [0,1]$ or $v \approx -u$.*

Any strict convex combination of u and v is more u-aligned than v. We also classify any expected-utility function as more u-aligned than $-u$, since $-u$ is maximally divergent from u.[11]

Theorem 1. *Suppose $(\succsim, \mathcal{C})$ has a Strotz representation $(u, v, \hat{v})$. Then the individual is naive if and only if $\hat{v} \gg_u v$ (and is sophisticated if and only if $\hat{v} \approx v$).*

Theorem 1 is a special case of the main result of the next section, where we turn to the more general case of random choice and uncertain beliefs.

3.2 General Results

In many environments, temptation is sensibly modeled as a random phenomenon. For example, someone might be motivated to work out at the gym on some days but lack enough willpower on other days. Even without temptation or naiveté, random choice provides a cleaner fit with noisy data in many applications. Uncertainty about future behavior is arguably more compelling when considering naiveté about temptation: Even if her actual future behavior is deterministic, a naive agent who cannot precisely predict her behavior might more naturally be modeled as having uncertainty about her future temptation, rather than making a resolute but incorrect prediction. Random temptation has been a part of many recent applications of time inconsistency and naiveté, ranging from optimal contracting (Eliaz and Spiegler (2006); Spiegler (2011)) to credit markets (Heidhues and Koszegi (2010)) to the design of commitment devices (Duflo, Kremer, and Robinson (2011)).

The conceptual apparatus just introduced for the deterministic case extends to random choice. For any (compound) lottery $\lambda^x \in \Delta(\Delta(C))$, its average choice $m(\lambda^x)$ is the expectation of the identity function under λ^x or, formally, $m(\lambda^x) = \int p \, d\lambda^x \in \Delta(C)$. That is, $m(\lambda^x)$ reduces the compound lottery λ^x into a single lottery in $\Delta(C)$. This reduction from a distribution over multiple lotteries to a single lottery does not assume any attitude towards risk, such as risk neutrality, over deterministic outcomes in C.[12]

[11] The special exception for this boundary case also has the technical benefit of avoiding tedious exceptions in the following characterization theorems.

[12] Our analysis does implicitly assume indifference to compounding. However, indifference to compounding can be relaxed by considering appropriate certainty equivalents rather than assuming indifference between λ^x and $m(\lambda^x)$.

Definition 4. *An individual is* sophisticated *if* $x \sim \{m(\lambda^x)\}$ *for all menus* x*. An individual is* naive *if* $x \succsim \{m(\lambda^x)\}$ *for all menus* x*. An individual is* strictly naive *if she is naive and not sophisticated.*

A sophisticate is indifferent between choosing from a menu x tomorrow and committing to the average choice $m(\lambda^x)$ from that menu. A naif anticipates making more virtuous choices, on average, than she actually will make. As noted above, deterministic second-stage choice formalized as a choice function $\mathcal{C} : \mathcal{K}(\Delta(C)) \to \Delta(C)$ is a special case of the random choice framework. The corresponding random choice rule λ satisfies

$$\lambda^x(\{p\}) = 1 \iff \mathcal{C}(x) = p,$$

and hence $m(\lambda^x) = \mathcal{C}(x)$. In this case our definitions of sophistication and naiveté reduce to $x \sim \{\mathcal{C}(x)\}$ and $x \succsim \{C(x)\}$, respectively.

Our definitions lend themselves to simple tests of violations of sophistication and naiveté. Consider a binary menu $\{p, q\}$ where $\{p\} \succ \{q\}$, and let $\alpha = \lambda^{\{p,q\}}(\{p\})$. Then, $m(\lambda^{\{p,q\}}) = \alpha p + (1 - \alpha)q$ and thus sophistication (naiveté) implies

$$\{p, q\} \sim (\succsim) \{\alpha p + (1 - \alpha)q\}.$$

In other words, a sophisticate is indifferent between the option set $\{p, q\}$ and a mixture of these lotteries that matches her ex-post choice frequencies, whereas a naif prefers keeping her options open. One possible experimental design that implements our approach would be to elicit the ranking of $\{p, q\}$ and $\{\hat{\alpha}p + (1 - \hat{\alpha})q\}$ for various values of $\hat{\alpha}$ and compare these rankings to the actual choice frequencies α of a group of subjects.[13]

We now apply our general definitions to the random Strotz model, which generalizes the classic Strotz model to allow uncertainty about future temptations. For example, a quasi-hyperbolic discounter may be uncertain of her degree of present bias. Dekel and Lipman (2012) provide a thorough analysis of the random Strotz model. Since a single temptation is parametrized as a single utility vector, a random temptation is analogously parametrized as a probability measure over utility vectors. Formally, let $\mathcal{V}$ denote the set of all continuous functions $v : C \to \mathbb{R}$. Endow $\mathcal{V}$ with the supremum norm and corresponding Borel σ-algebra. We can identify $\mathcal{V}$ with the set of all expected-utility functions on $\Delta(C)$ by letting $v(p) \equiv \int_C v(c)\, dp$.

Definition 5. *A probability measure* μ *on* $\mathcal{V}$ *has* finite-dimensional support *if there exists a finite set of expected-utility functions* $\{v_1, \ldots, v_n\} \subset \mathcal{V}$ *such that* $\text{supp}(\mu) \subset \text{span}(\{v_1, \ldots, v_n\})$.

[13] This design is implemented in Le Yaouanq (2015) to measure individual-level naiveté about memory lapses.

We restrict attention to random Strotz representations with finite-dimensional support. This entails no practical limitations; we are unaware of any application of the random Strotz model without finite-dimensional support. For example, any deterministic Strotz representation (see Definition 2) or any uncertain intensity random Strotz representation (see Appendix B) such as random quasi-hyperbolic discounting (see Section 5.1) has finite-dimensional support. In addition, if the consumption space C is finite, then any probability measure μ on $\mathcal{V}$ trivially has finite-dimensional support.

Without loss of generality, we also restrict attention to probability measures on $\mathcal{V}$ that are *nontrivial*, in the sense of assigning probability zero to constant functions.[14]

Definition 6. *A* random Strotz representation *of* $(\succsim, \lambda)$ *is a triple* $(u, \mu, \hat{\mu})$ *of a nontrivial expected-utility function* u *and nontrivial probability measures* μ *and* $\hat{\mu}$ *over* $\mathcal{V}$ *with finite-dimensional support such that the function* $U : \mathcal{K}(\Delta(C)) \to \mathbb{R}$ *defined by*

$$U(x) = \int_{\mathcal{V}} \max_{p \in B_v(x)} u(p) \, d\hat{\mu}(v)$$

is a utility representation of $\succsim$ *and, for all menus* x *and all measurable* $y \subset x$*,*

$$\lambda^x(y) = \mu(p_x^{-1}(y))$$

for some measurable selection function $p_x : \mathcal{V} \to x$ *with* $p_x(v) \in B_u(B_v(x))$ *for all* $v \in \mathcal{V}$*.*[15]

The interpretation of the representation of the ex-ante preference $\succsim$ is straightforward. To understand the representation of the ex-post random choice rule λ, note that after the realization of a temptation utility $v \in \mathcal{V}$, the individual's choice of lottery is an element of the set $B_u(B_v(x))$ of lexicographic maximizers of v then u. There may be multiple elements in this set for a fixed v, and the individual's tie-breaking procedure among these is modeled using a selection function p_x from the correspondence $v \mapsto B_u(B_v(x))$ mapping temptations to possible choices.[16] Given this mapping from temptation utilities to choices, the distribution of temptation utilities then determines the stochastic choice

[14]The restriction to nontrivial measures in the definition of the random Strotz representation is also without loss of generality since any weight assigned to constant functions can be moved to the commitment utility u without altering the ex-ante preference or ex-post random choice rule.

[15]Note that Definition 2 is equivalent to the special case of Definition 6 where $\mu = \delta_v$ and $\hat{\mu} = \delta_{\hat{v}}$ for some fixed $v, \hat{v}$. The latter implies $\lambda^x(\{p_x(v)\}) = 1$ in this case or, equivalently, $\mathcal{C}(x) = p_x(v) \in B_u(B_v(x))$.

[16]Since there may be a multiplicity of selection functions, there may in turn be multiple maximizing choice probabilities over x for a fixed probability measure μ over $\mathcal{V}$. That is, just as there can be a multiple choice *functions* induced by a choice *correspondence*, there can be multiple random choice rules that maximize the same random Strotz representation. However, this multiplicity is not important for our results since observing any maximizing random choice rule provides sufficient information for our comparatives.

of the individual. The probability of choosing an element of the subset $y \subset x$ is equal to the probability under μ of an ex-post expected-utility function v for which the optimal choice is in y, $\lambda^x(y) = \mu(\{v \in \mathcal{V} : p_x(v) \in y\})$.

The definition of naiveté for random Strotz is the stochastic generalization of the definition for deterministic Strotz. In the degenerate case, naiveté implies the believed $\hat{v}$ is more u-aligned than v: $\hat{v} \gg_u v$. In the random case, the believed distribution over all possible temptations stochastically dominates the actual distribution of temptations, where stochastic dominance is with respect to the $\gg_u$ order. As is standard, a stochastically dominant measure puts more weight on the upper contour sets of the basic ordering $\gg_u$ over the state space. The following definitions adapt the technology developed by Dekel and Lipman (2012).

Definition 7. *Let u be an expected-utility function. A measurable set $\mathcal{U} \subset \mathcal{V}$ is a* u-upper set *if, for any $v \in \mathcal{U}$ and $v' \in \mathcal{V}$, if $v' \gg_u v$ then $v' \in \mathcal{U}$.*

We let $\gg_u$ notate both the basic ordering over expected-utility functions and the induced stochastic order over measures on expected-utility functions.

Definition 8. *Let u be an expected-utility function, and let $\mu, \hat{\mu}$ be probability measures over $\mathcal{V}$. Then $\hat{\mu}$ is* more u-aligned *than μ, written as $\hat{\mu} \gg_u \mu$, if $\hat{\mu}(\mathcal{U}) \geq \mu(\mathcal{U})$ for all u-upper sets $\mathcal{U}$.*

Note that $\hat{v} \gg_u v$ (in the determinate sense) is equivalent to $\delta_{\hat{v}} \gg_u \delta_v$ (in the stochastic sense). We write $\hat{\mu} \approx \mu$ whenever both $\hat{\mu} \gg_u \mu$ and $\mu \gg_u \hat{\mu}$, that is, when $\hat{\mu}(\mathcal{U}) = \mu(\mathcal{U})$ for all u-upper sets $\mathcal{U}$. In this case, it can be shown that the measures induce identical distributions over ex-post expected-utility *preferences* and can differ only by affine transformations of the utility functions in their supports.[17] They are therefore identical in every respect that is relevant for both ex-ante and ex-post choice.

Generalizing our earlier result, absolute naiveté is equivalent to $\hat{\mu}$ dominating μ in the stochastic order generated by $\gg_u$.

Theorem 2. *Suppose $(\succsim, \lambda)$ has a random Strotz representation $(u, \mu, \hat{\mu})$. Then the individual is naive if and only if $\hat{\mu} \gg_u \mu$ (and is sophisticated if and only if $\hat{\mu} \approx \mu$).*

The proof of this result makes use of a characterization by Dekel and Lipman (2012) of comparative temptation aversion for ex-ante preferences with random Strotz representations. They say that $\succsim_2$ is *more temptation averse* than $\succsim_1$ if, for all menus x and

[17] The formal statement and proof of this claim can be found in Dekel and Lipman (2012); in particular, see their Theorem 3 and its proof.

lotteries p,[18]

$$\{p\} \succ_1 x \implies \{p\} \succ_2 x.$$

Dekel and Lipman (2012) show that if $\succsim_i$ has a random Strotz representation (u, μ_i) for $i = 1, 2$, then $\succsim_2$ is more temptation averse than $\succsim_1$ if and only if $\mu_1 \gg_u \mu_2$. To prove Theorem 2, we apply this comparative to the measures $\hat{\mu}$ and μ in our two-period random Strotz representation for a single individual. In particular, we show that naiveté is equivalent to the condition

$$\int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\hat{\mu}(v) = U(x) \geq u(m(\lambda^x)) = \int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\mu(v), \quad \forall x.$$

This condition implies that the hypothetical ex-ante preference $\succsim^*$ generated by the representation with correct beliefs (u, μ) is more temptation averse than the actual ex-ante preference $\succsim$ with representation $(u, \hat{\mu})$, and hence $\hat{\mu} \gg_u \mu$.

Two special cases of the random Strotz representation will be useful for illustrating the conditions in this theorem, as well as our subsequent results on comparative naiveté. The first is the deterministic Strotz representation already described in Definition 2: Theorem 1 follows as a corollary of Theorem 2 by taking $\hat{\mu} = \delta_{\hat{v}}$ and $\mu = \delta_v$. The second is a simple stochastic model proposed by Eliaz and Spiegler (2006) in which the individual has temptation utility v with probability $1 - \theta$ and no temptation with probability θ. We say that two expected-utility functions u and v are *independent* if they are nontrivial and it is not the case that $v \approx u$ or $v \approx -u$.

Definition 9. *An* Eliaz-Spiegler representation *of* $(\succsim, \lambda)$ *is a quadruple* $(u, v, \theta, \hat{\theta})$ *of independent expected-utility functions* u *and* v *and scalars* $\theta, \hat{\theta} \in [0, 1]$ *such that the function* $U : \mathcal{K}(\Delta(C)) \to \mathbb{R}$ *defined by*

$$U(x) = \hat{\theta} \max_{p \in x} u(p) + (1 - \hat{\theta}) \max_{p \in B_v(x)} u(p)$$

is a utility representation of $\succsim$ *and, for all menus* x*,*

$$\lambda^x = \theta \delta_{p_u} + (1 - \theta) \delta_{p_v}$$

for some $p_u \in B_u(x)$ *and* $p_v \in B_u(B_v(x))$*.*

Theorem 2 yields the following corollary by taking $\mu = \theta\delta_u + (1 - \theta)\delta_v$ and $\hat{\mu} = \hat{\theta}\delta_u + (1 - \hat{\theta})\delta_v$ and observing that $\mu(\mathcal{U}) = \theta$ and $\hat{\mu}(\mathcal{U}) = \hat{\theta}$ for the u-upper set $\mathcal{U}$ containing only the positive affine transformations of u.

[18] This formal definition appears with different interpretations in Ahn (2007) and Sarver (2008). It is also similar in spirit to the behavioral comparisons of ambiguity aversion in Epstein (1999) and Ghirardato and Marinacci (2002), who compare arbitrary acts to unambiguous acts in the same manner that we compare arbitrary menus to singleton menus.

Corollary 1. *Suppose $(\succsim, \lambda)$ has an Eliaz-Spiegler representation $(u, v, \theta, \hat{\theta})$. Then the individual is naive if and only if $\hat{\theta} \geq \theta$ (and is sophisticated if and only if $\hat{\theta} = \theta$).*

4 Comparisons of Naiveté

In this section, we introduce two definitions for comparing naiveté across agents. The first naturally extends our proposed test for absolute naiveté by counting passed opportunities for beneficial commitment; the second directly measures the difference in anticipated and actual indirect utilities for menus. As with the case of risk aversion, different comparisons can be useful depending on the application at hand.

4.1 A Strong Comparison of Naiveté

Having proposed a behavioral definition of absolute naiveté, we naturally consider the comparison of niaveté across heterogeneous individuals. Recall that a naive agent satisfies $x \succsim \{m(\lambda^x)\}$, that is, there is a potential gap between her value for the the menu x above her eventual expected choice $m(\lambda^x)$. To compare the degree of naiveté across agents, we propose measuring the size of this gap through preference for commitment.

Definition 10. *Individual 1 is* more naive *than individual 2 if, for all menus x and lotteries p,*

$$x \succsim_2 \{p\} \succsim_2 \{m(\lambda_2^x)\} \implies x \succsim_1 \{p\} \succsim_1 \{m(\lambda_1^x)\}.$$

Any commitment p that is ex-ante ranked between x and $\{m(\lambda^x)\}$ indicates naiveté, because p is more virtuous than the expected choice $m(\lambda^x)$ yet the individual prefers to maintain the flexibility in x. So, the welfare-improving opportunity to commit to p will be naively rejected. If another individual is more naive, then she would also reject that commitment.[19]

In the random Strotz model, this definition imposes sharp and intuitive restrictions on the believed and actual temptations of both agents.

Theorem 3. *Suppose $(\succsim_1, \lambda_1)$ and $(\succsim_2, \lambda_2)$ are naive and have random Strotz representations $(u, \mu_1, \hat{\mu}_1)$ and $(u, \mu_2, \hat{\mu}_2)$. Then individual 1 is more naive than individual 2 if and only if*

$$\hat{\mu}_1 \gg_u \hat{\mu}_2 \gg_u \mu_2 \gg_u \mu_1.$$

[19] An alternative formulation of more naive based on strict preferences is also possible:

$$x \succ_2 \{p\} \succ_2 \{m(\lambda_2^x)\} \implies x \succ_1 \{p\} \succ_1 \{m(\lambda_1^x)\}.$$

Our results will carry over to this case, as long as individual 2 is strictly naive (the condition is vacuously satisfied if individual 2 is sophisticated).

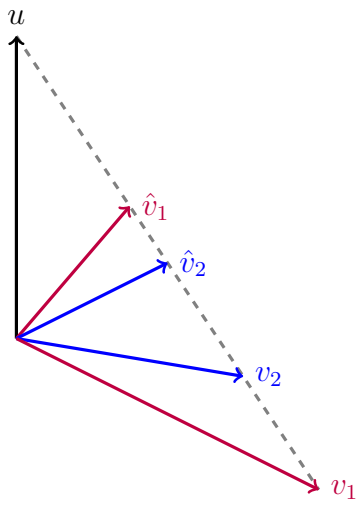

Figure 1: Alignment of believed and actual utilities implied by comparative naiveté in the (deterministic) Strotz representation (Corollary 2).

While they share common normative preferences over singleton commitments, individual 1 is more optimistic about her future behavior than individual 2, as reflected in the requirement $\hat{\mu}_1 \gg_u \hat{\mu}_2$. However, individual 1's actual ex-post choices are even less virtuous than individual 2's choices, as reflected in $\mu_2 \gg_u \mu_1$. A more naive individual is more optimistic about her future virtuous behavior while actually exercising less virtue.

Taking $\mu_i = \delta_{v_i}$ and $\hat{\mu}_i = \delta_{\hat{v}_i}$ in Theorem 3 yields the following corollary for deterministic Strotz representations.

Corollary 2. *Suppose* $(\succsim_1, \mathcal{C}_1)$ *and* $(\succsim_2, \mathcal{C}_2)$ *are naive and have Strotz representations* $(u, v_1, \hat{v}_1)$ *and* $(u, v_2, \hat{v}_2)$*. Then individual* 1 *is more naive than individual* 2 *if and only if*

$$\hat{v}_1 \gg_u \hat{v}_2 \gg_u v_2 \gg_u v_1.$$

As illustrated in Figure 1, comparative naiveté implies that both individuals' anticipated temptations $\hat{v}_i$ and actual temptations v_i are convex combinations of the shared commitment utility u and the more naive individual's actual temptation v_1, progressively located on the arc connecting u and v_1.

Theorem 3 yields the following corollary for the Eliaz-Spiegler representation by taking $\mu_i = \theta_i \delta_u + (1 - \theta_i)\delta_v$ and $\hat{\mu}_i = \hat{\theta}_i \delta_u + (1 - \hat{\theta}_i)\delta_v$.

Corollary 3. *Suppose* $(\succsim_1, \lambda_1)$ *and* $(\succsim_2, \lambda_2)$ *are naive and have Eliaz-Spiegler representations* $(u, v, \theta_1, \hat{\theta}_1)$ *and* $(u, v, \theta_2, \hat{\theta}_2)$*. Then individual* 1 *is more naive than individual* 2 *if and only if* $\hat{\theta}_1 \geq \hat{\theta}_2 \geq \theta_2 \geq \theta_1$.[20]

[20]This result can also be extended to Eliaz-Spiegler representations $(u, v_1, \theta_1, \hat{\theta}_1)$ and $(u, v_2, \theta_2, \hat{\theta}_2)$

4.2 Quantitative Measures and a More Complete Ordering

In many applications of time inconsistency and naiveté to industrial organization and contract theory, the firm's ability to extract excess surplus is tied to the extent to which the individual overestimates the utility that she will receive from a set of options or contract.[21] This motivates the following construction.

Definition 11. *Suppose* $(\succsim, \lambda)$ *has a random Strotz representation* $(u, \mu, \hat{\mu})$. *The* coefficient of over-valuation *of a menu* x *is defined by:*

$$OV(x) = \underbrace{\int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\hat{\mu}(v)}_{\text{believed indirect utility}} - \underbrace{\int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\mu(v)}_{\text{actual indirect utility}}.$$

A natural conjecture is that our prior definition of more naive is equivalent to having a higher over-valuation for every menu x. This is false: Our behavioral comparative is sufficient but not necessary. Individual 1 is more naive than individual 2 if and only if, for every menu x,[22]

$$\begin{aligned} \underbrace{\int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\hat{\mu}_1(v)}_{\text{1's believed indirect utility}} &\geq \underbrace{\int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\hat{\mu}_2(v)}_{\text{2's believed indirect utility}} \\ &\geq \underbrace{\int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\mu_2(v)}_{\text{2's actual indirect utility}} \geq \underbrace{\int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\mu_1(v)}_{\text{1's actual indirect utility}}. \end{aligned} \tag{1}$$

The inequalities in Equation (1) exclude some cases where individual 1 is more susceptible to exploitation than individual 2. For example, suppose two individuals' random temptations μ_1 and μ_2 are not comparable under the $\gg_u$ order, but individual 2 is sophisticated while individual 1 is strictly naive. Individual 1 is not more naive than individual 2 according to Definition 10 because their ex-post behaviors are not $\gg_u$-ranked, but the over-valuation of any menu x is higher for 1 than for 2, $OV_1(x) \geq OV_2(x) = 0$, with strict inequality for some menu.

In this section, we examine the weaker comparison of naiveté based on over-valuation and establish its equivalence to several other behavioral and quantitative measures. As

where $v_1 \neq v_2$ is permitted. In this case, individual 1 is more naive than individual 2 if and only if $\hat{\theta}_1 \geq \hat{\theta}_2 \geq \theta_2 \geq \theta_1$ and, in addition, $v_2 \gg_u v_1$ whenever $\theta_2 < 1$ and $v_2 \approx v_1$ whenever $\hat{\theta}_1 < 1$.

[21] Some applications are reviewed in Spiegler (2011) and Koszegi (2014, Section 6).

[22] It is easy to verify that individual 1 is more naive than individual 2 if and only if $U_1(x) \geq U_2(x) \geq u(m(\lambda_2^x)) \geq u(m(\lambda_1^x))$ for every menu x. Using Lemma 2 in Appendix C.1, this condition is equivalent to Equation (1).

just mentioned, having higher over-valuation is weaker than our previous definition of more naive, so a correspondingly less stringent behavioral comparative is needed.

Our model incorporates all relevant dimensions of consumption—potentially including goods, effort, and money—into the space C. But for the sake of developing intuition for how to calibrate over-valuation from choice data, consider a special quasilinear environment where ex-ante choices are over pairs of a menu $x \in \mathcal{K}(\Delta(C))$ and a money transfer $t \in \mathbb{R}$, and ex-ante utility takes the form $V(x,t) = U(x) + t$. By its definition, the over-valuation of the menu x must satisfy

$$(x, 0) \sim (m(\lambda^x), OV(x)).$$

The required monetary premium for x relative to $m(\lambda^x)$ immediately quantifies over-valuation for quasilinear preferences. Then an immediate behavioral comparative is that individual 1 is willing to overpay more for any menu x than individual 2:

$$(x, 0) \succsim_2 (m(\lambda_2^x), t) \implies (x, 0) \succsim_1 (m(\lambda_1^x), t).$$

This condition is equivalent to $OV_1(x) \geq OV_2(x)$.

Since our general model does not assume quasilinearity, we must take a different approach to calibrating over-valuation. As a side benefit of assuming expected utility, we can replace the numeraire with linearity in probabilities to measure the value of x relative to $m(\lambda_i^x)$. The following definition takes this approach to converting over-valuation into a behavioral measure.

Definition 12. *Fix any lotteries p, q such that $\{q\} \succ \{p\}$. The* probability premium *of a menu x is defined by:*

$$P(x; p, q) = \sup \big\{\alpha \in [0,1] : (1-\alpha)x + \alpha\{p\} \succsim (1-\alpha)\{m(\lambda^x)\} + \alpha\{q\}\big\}.$$

The probability premium indicates how much a menu x can be mixed with an inferior alternative with the individual still preferring it to $m(\lambda^x)$ mixed with a superior alternative. To see its implications, suppose that $\succsim$ admits an affine utility representation. Then note that $P(x; p, q) < 1$ for any menu x since $\{q\} \succ \{p\}$, and $P(x; p, q) = 0$ if and only if $x \sim \{m(\lambda^x)\}$. In particular, the individual is sophisticated if and only if $P(x; p, q) = 0$ for all x.

The following definition defines another behavioral comparative of naiveté based on a similar approach. The definition exploits the separability across events afforded by the independence axiom to use the outcomes in other events (the half-probability events that p or q are the relevant lottery) to measure the value of x relative to $m(\lambda_i^x)$.

Definition 13. *Individual 1 is* weakly more naive *than individual 2 if, for all menus x and lotteries p, q,*

$$\frac{1}{2}x + \frac{1}{2}\{p\} \succsim_2 \frac{1}{2}\{m(\lambda_2^x)\} + \frac{1}{2}\{q\} \implies \frac{1}{2}x + \frac{1}{2}\{p\} \succsim_1 \frac{1}{2}\{m(\lambda_1^x)\} + \frac{1}{2}\{q\}.$$

As the name suggests, this comparative measure is indeed weaker than our previous definition of more naive. Under mild technical assumptions, Definition 13 is implied by Definition 10, and hence this is a more complete ordering of individuals.

Lemma 1. *Suppose $\succsim_1$ and $\succsim_2$ satisfy independence and share the same commitment preference.*[23] *Suppose also that individual 2 is naive and that, for all menus x, there exists a lottery p such that $x \sim_2 \{p\}$. If individual 1 is more naive than individual 2, then 1 is weakly more naive than 2.*

While more permissive than the first definition of naiveté, the weak definition still yields several useful equivalent characterizations for random Strotz preferences, unifying comparisons of naiveté based on probability premia, over-valuations of contracts, and stochastic ordering of the *differences* between believed and actual random temptations.

Theorem 4. *Suppose $(\succsim_1, \lambda_1)$ and $(\succsim_2, \lambda_2)$ are naive and have random Strotz representations $(u, \mu_1, \hat{\mu}_1)$ and $(u, \mu_2, \hat{\mu}_2)$. Fixing any lotteries p, q with $\{q\} \succ_i \{p\}$, the following are equivalent:*

1. *Individual 1 is weakly more naive than individual 2.*
2. *$P_1(x; p, q) \geq P_2(x; p, q)$ for all menus x.*
3. *$OV_1(x) \geq OV_2(x)$ for all menus x.*
4. *$\hat{\mu}_1(\mathcal{U}) - \mu_1(\mathcal{U}) \geq \hat{\mu}_2(\mathcal{U}) - \mu_2(\mathcal{U})$ for all u-upper sets $\mathcal{U}$; equivalently, $\hat{\mu}_1 - \mu_1 \gg_u \hat{\mu}_2 - \mu_2$.*

Specializing condition (4) to Eliaz-Spiegler preferences yields the following comparison that ranks naiveté by the difference in the believed and actual probabilities of being virtuous.

Corollary 4. *Suppose $(\succsim_1, \lambda_1)$ and $(\succsim_2, \lambda_2)$ are naive and have Eliaz-Spiegler representations $(u, v, \theta_1, \hat{\theta}_1)$ and $(u, v, \theta_2, \hat{\theta}_2)$. Then individual 1 is weakly more naive than individual 2 if and only if $\hat{\theta}_1 - \theta_1 \geq \hat{\theta}_2 - \theta_2$.*[24]

[23] The preference $\succsim_i$ satisfies independence if for any menus x, y, z and $\alpha \in (0, 1)$, $x \succsim_i y$ implies $\alpha x + (1-\alpha)z \succsim_i \alpha y + (1-\alpha)z$. We say $\succsim_1$ and $\succsim_2$ share the same commitment preference if for all lotteries p, q, $\{p\} \succsim_1 \{q\} \iff \{p\} \succsim_2 \{q\}$.

[24] This result can also be extended to Eliaz-Spiegler representations $(u, v_1, \theta_1, \hat{\theta}_1)$ and $(u, v_2, \theta_2, \hat{\theta}_2)$ where $v_1 \neq v_2$ is permitted. In this case, individual 1 is weakly more naive than individual 2 if and only if $\hat{\theta}_1 - \theta_1 \geq \hat{\theta}_2 - \theta_2$ and, in addition, $v_2 \gg_u v_1$ whenever $\hat{\theta}_2 > \theta_2$ (individual 2 is strictly naive).

While the two comparisons are generally different for random choices, there is one prominent case where the weak and strong notions of comparative naiveté align: deterministic Strotz. The next result follows from considering the special case of condition (4) in Theorem 4 where each measure is a deterministic point mass.

Corollary 5. *Suppose* $(\succsim_1, \mathcal{C}_1)$ *and* $(\succsim_2, \mathcal{C}_2)$ *are naive and have Strotz representations* $(u, v_1, \hat{v}_1)$ *and* $(u, v_2, \hat{v}_2)$. *Then individual* 1 *is weakly more naive than individual* 2 *if and only if either*

$$\hat{v}_1 \gg_u \hat{v}_2 \gg_u v_2 \gg_u v_1$$

or $\hat{v}_2 \approx v_2$ *(individual 2 is sophisticated).*

Comparing differences in probability parameters has a behavioral justification for random choice, as evidenced in the implications for Eliaz-Spiegler preferences. But comparing differences in parameters is not sensible for deterministic models. For example, we will show in Section 5.1 that a consequence of Corollary 5 for the quasi-hyperbolic discounting model is that ranking naiveté by the restriction $\hat{\beta}_1 - \beta_1 \geq \hat{\beta}_2 - \beta_2$ does not correspond to either definition of comparative naiveté.

5 Applications to Present Bias

To illustrate the general appropriateness of our definitions, we consider their implications for two models of present bias that generalize the ubiquitous quasi-hyperbolic discounting model. In Section 5.1, we apply our results to a stochastic generalization of quasi-hyperbolic discounting that permits uncertainty about the degree of time inconsistency. In Section 5.2, we analyze a deterministic model of present bias that permits more general patterns of time discounting, such as true hyperbolic discounting. The results in these sections show that our definitions of absolute and comparative naiveté not only confirm known parametric formulations of naiveté (as in the special case of deterministic quasi-hyperbolic discounting), but also generate new insights for other well-known models for which comparisons of naiveté are still outstanding.

5.1 Random Quasi-Hyperbolic Discounting

As a specific application of the previous characterizations, we consider the random quasi-hyperbolic model in which time inconsistency is parameterized by a present-bias factor β and the individual may be uncertain about the value for this parameter. Let $C = [a, b]^{\mathbb{N}}$ be a set of infinite-horizon consumption streams, with elements $c = (c_1, c_2, \dots) \in C$.[25]

[25] The product topology on C is compact and metrizable.

A lottery $p \in \Delta(C)$ resolves immediately and yields a consumption stream. We focus on the simple case with one-shot resolution of uncertainty for expositional parsimony, but all of the following results generalize to richer settings that incorporate temporal lotteries or true dynamic choice.[26] In these more general dynamic environments, simple atemporal lotteries over consumption streams provide sufficient choice observations to apply the following comparative statics.

Suppose the commitment preference is represented by an expected-utility function whose values $u(c) = u(\delta_c)$ over deterministic streams (that is, whose Bernoulli utility indices) comply with exponential discounting,

$$u(c) = \sum_{t=1}^{\infty} \delta^{t-1} w(c_t), \tag{2}$$

for some instantaneous utility function $w : [a, b] \to \mathbb{R}$. The quasi-hyperbolic discounting model captures present bias with an additional discount factor applied to all future periods: If the present-bias factor is β, then ex-post (period 1) choice from a menu of consumption streams x will maximize

$$v_\beta(c) = w(c_1) + \beta \sum_{t=2}^{\infty} \delta^{t-1} w(c_t). \tag{3}$$

We begin by defining the quasi-hyperbolic discounting representation for deterministic choice before proceeding to its stochastic generalization. In the deterministic model, the individual's ex-ante (period 0) behavior may reflect an incorrect belief that her future present-bias parameter is $\hat{\beta}$, while her ex-post behavior actually uses the present-bias parameter β. It is immediate that this choice procedure corresponds to a special case of the deterministic Strotz representation.

Definition 14. *A* quasi-hyperbolic (QH) representation *of* $(\succsim, \mathcal{C})$ *is a tuple* $(w, \beta, \hat{\beta}, \delta)$ *of a continuous and nontrivial function* $w : [a, b] \to \mathbb{R}$ *and scalars* $\beta, \hat{\beta} \in [0, 1]$ *and* $\delta \in (0, 1)$*, such that* $(u, v_\beta, v_{\hat{\beta}})$ *defined as in Equations* (2) *and* (3) *for these parameters is a Strotz representation for* $(\succsim, \mathcal{C})$*.*

The standard quasi-hyperbolic discounting model assumes completely confident beliefs about future behavior, an assumption that seems less palatable under naiveté when these beliefs are incorrect. We explore a generalization of the QH representation that allows for naive and uncertain beliefs about β. Several applications in different areas employ naive uncertainty about future present bias. Heidhues and Koszegi (2010, Section 4)

[26] Kreps and Porteus (1978) were the first to provide a complete analysis of dynamic choice with uncertainty that resolves gradually through time (i.e., temporal lotteries). The models of temptation in Gul and Pesendorfer (2004) and Noor (2011) used an infinite-horizon version of such a setting.

employ random quasi-hyperbolic discounting to explain the structure of credit markets and its consequent welfare implications. In their study of fertilizer adoption decisions by Kenyan farmers, Duflo, Kremer, and Robinson (2011) estimate a specification of random quasi-hyperbolic discounting where naiveté is parameterized by a mistakenly believed positive chance of virtuous exponential discounting. Admitting uncertainty about intertemporal substitution often usefully serves as a reduced-form proxy for a shock in the economy, like wage uncertainty, or for heterogeneity across agents in an aggregate economy, like the distribution of wealth. Similarly, random present-bias can provide a parsimonious channel for capturing uncertainty about external factors that affect present-bias.

Definition 15. *A* random quasi-hyperbolic (RQH) representation *of* $(\succsim, \lambda)$ *is a quadruple* $(w, F, \hat{F}, \delta)$ *of a continuous and nontrivial function* $w : [a, b] \to \mathbb{R}$*, a scalar* $\delta \in (0, 1)$*, and cumulative distribution functions* F *and* $\hat{F}$ *on* $[0, 1]$ *such that when* u *and* v_β *are defined as in Equations* (2) *and* (3)*, the function* $U : \mathcal{K}(\Delta(C)) \to \mathbb{R}$ *defined by*

$$U(x) = \int_0^1 \max_{p \in B_{v_\beta}(x)} u(p)\, d\hat{F}(\beta)$$

is a utility representation of $\succsim$ *and, for all menus* x *and all measurable* $y \subset x$*,*

$$\lambda^x(y) = F(p_x^{-1}(y))$$

for some measurable selection function $p_x : [0, 1] \to x$ *with* $p_x(\beta) \in B_u(B_{v_\beta}(x))$ *for all* $\beta \in [0, 1]$*.*[27]

The RQH representation is a member of a more general subclass of the random Strotz representation where the possible temptations are ordered by a one-dimensional parameter. We analyze this subclass, called the uncertain intensity Strotz representation, in Appendix B. The corollaries presented below follow directly from the results in that section.

A naive individual underestimates the degree of her present bias, which is reflected in her belief $\hat{F}$ putting more likelihood on larger values of β than the actual distribution F that governs her ex-post choices. Let $\geq_{FOSD}$ denote the usual first-order stochastic dominance order, with $\hat{F} \geq_{FOSD} F$ if $\hat{F}(\beta) \leq F(\beta)$ for all $\beta \in [0, 1]$.

Corollary 6. *Suppose* $(\succsim, \lambda)$ *has a RQH representation* $(w, F, \hat{F}, \delta)$*. Then the individual is naive if and only if* $\hat{F} \geq_{FOSD} F$ *(and is sophisticated if and only if* $\hat{F} = F$*).*

[27] We are abusing notation slightly and using F to also denote the probability measure on $[0, 1]$ that has F as its distribution function. That is, for any measurable set $A \subset [0, 1]$, we write $F(A)$ to denote $\int_A dF(\beta)$. Hence $\lambda^x(y) = \int_0^1 \mathbf{1}_{[p_x(\beta) \in y]}\, dF(\beta)$.

Corollary 7. *Suppose* $(\succsim_1, \lambda_1)$ *and* $(\succsim_2, \lambda_2)$ *are naive and have RQH representations* $(w, F_1, \hat{F}_1, \delta)$ *and* $(w, F_2, \hat{F}_2, \delta)$.

1. *Individual 1 is more naive than individual 2 if and only if*

$$\hat{F}_1 \geq_{FOSD} \hat{F}_2 \geq_{FOSD} F_2 \geq_{FOSD} F_1.$$

2. *Individual 1 is weakly more naive than individual 2 if and only if*

$$F_1(\beta) - \hat{F}_1(\beta) \geq F_2(\beta) - \hat{F}_2(\beta), \quad \forall \beta \in [0, 1].$$

In the case of deterministic quasi-hyperbolic discounting, both of our comparative measures collapse to the same condition, excepting the special case where individual 2 is sophisticated.

Corollary 8. *Suppose* $(\succsim_1, \mathcal{C}_1)$ *and* $(\succsim_2, \mathcal{C}_2)$ *are naive and have QH representations* $(w, \beta_1, \hat{\beta}_1, \delta)$ *and* $(w, \beta_2, \hat{\beta}_2, \delta)$.

1. *Individual 1 is more naive than individual 2 if and only if* $\hat{\beta}_1 \geq \hat{\beta}_2 \geq \beta_2 \geq \beta_1$.
2. *Individual* 1 *is weakly more naive than individual* 2 *if and only if either* $\hat{\beta}_1 \geq \hat{\beta}_2 \geq \beta_2 \geq \beta_1$ *or* $\hat{\beta}_2 = \beta_2$ *(individual 2 is sophisticated).*

Corollary 8 provides another set of intuitive comparative restrictions. First, the more naive individual has more optimistic beliefs about her future patience: $\hat{\beta}_1 \geq \hat{\beta}_2$. Second, the more naive individual's behavior is more present-biased: $\beta_1 \leq \beta_2$. In contrast to Eliaz and Spiegler (2006) preferences (cf. Corollaries 3 and 4), the weaker second comparison of naiveté does not greatly generalize the strong definition, adding only comparisons with purely sophisticated agents. For example, while $\hat{\beta}_1 - \beta_1 \geq \hat{\beta}_2 - \beta_2$ may seem like an appealing comparison of naiveté, this inequality alone is generally insufficient to guarantee individual 1 is more naive than individual 2.

5.2 Diminishing Impatience

The prior analysis of the quasi-hyperbolic representation extends to more general patterns of discounting, such as true hyperbolic discounting. We now relate several properties of discount functions to properties of the perceived discount functions for individuals who satisfy our definition of naiveté. While our definition corroborates the existing parameter

restriction $\hat{\beta} \geq \beta$ for naiveté with deterministic quasi-hyerbolic discounting, the analogous formulation for general diminishing impatience is less understood.[28] This section introduces the appropriate restrictions and uncovers structural relationships between the underestimation of impatience and actual impatience that declines over time.

Say that $D : \mathbb{N} \cup \{0\} \to (0, 1]$ is a *discount function* if $D(0) = 1$ and $\sum_{t=0}^{\infty} D(t) < \infty$. Suppose as before that consumption in periods $t = 1, 2, \dots$ is given by $(c_1, c_2, \dots) \in C = [a, b]^{\mathbb{N}}$. Period 0 commitment preferences over deterministic consumption streams starting in period 1 are represented by

$$u(c) = \sum_{t=1}^{\infty} D(t)w(c_t). \tag{4}$$

Suppose that preferences over consumption streams are stationary, so period 1 choices maximize

$$v(c) = \sum_{t=1}^{\infty} D(t-1)w(c_t). \tag{5}$$

However, in period 0 the individual believes that she will apply the discount function $\hat{D}$ in the subsequent period, which yields the following anticipated temptation utility for deterministic consumption streams:

$$\hat{v}(c) = \sum_{t=1}^{\infty} \hat{D}(t-1)w(c_t). \tag{6}$$

Definition 16. *A* discounting representation *of* $(\succsim, \mathcal{C})$ *is a triple* $(w, D, \hat{D})$ *of a continuous and nontrivial function* $w : [a, b] \to \mathbb{R}$ *and discount functions* D *and* $\hat{D}$*, such that* $(u, v, \hat{v})$ *defined as in Equations* (4)*,* (5)*, and* (6) *is a Strotz representation for* $(\succsim, \mathcal{C})$*.*

The deterministic quasi-hyperbolic representations discussed in the previous section are special cases of the discounting representations where

$$D(t) = \begin{cases} 1 & \text{if } t = 0 \\ \beta\delta^t & \text{if } t > 0. \end{cases}$$

and

$$\hat{D}(t) = \begin{cases} 1 & \text{if } t = 0 \\ \hat{\beta}\delta^t & \text{if } t > 0. \end{cases}$$

[28] Prelec (2004) studies the degree of time inconsistency for a single discount function D, as captured by log-concavity. He suggests this as a criterion for evaluating sophistication, but this approach is clearly conceptually remote from our notion of sophistication that relies on comparing D with a believed discount funtion $\hat{D}$.

Two general properties of discount functions will be important.

Definition 17. *A discount function* $D : \mathbb{N} \cup \{0\} \to (0, 1]$ *exhibits* diminishing impatience *if*

$$\frac{D(0)}{D(1)} > \frac{D(t)}{D(t+1)} \quad (\forall t \in \mathbb{N}),$$

and exhibits strong diminishing impatience *if*

$$\frac{D(t)}{D(t+1)} > \frac{D(t+1)}{D(t+2)} \quad (\forall t \in \mathbb{N} \cup \{0\}).$$

Diminishing impatience requires that the discount rate for any pair of successive periods in the future is strictly more balanced than the discount rate between today and tomorrow. Strong diminishing impatience further requires that the discount rate between successive periods is strictly declining over time. Quasi-hyperbolic discount functions exhibit diminishing impatience but not strong diminishing impatience because the discount rate between t and $t+1$ is constant at $1/\delta$ after $t = 1$, whereas true hyperbolic discounting, on the other hand, exhibits strong diminishing impatience.

The following corollary of Theorem 1 uncovers the implications of diminishing and strong diminishing impatience on the perceived future impatience of a naive individual. The individual believes that her ex-post intertemporal rate of substitution between period 1 and period $t+1$ will be governed by the discount factor $\hat{D}(t)$. This discount factor is a convex combination of the ex-ante discount factor $D(t+1)/D(1)$ that would be applied if committing in period 0 and the actual tempting discount factor $D(t)$ that governs the intertemporal consumption stream that the individual will actually choose tomorrow.

Corollary 9. *Suppose* $(\succsim, \mathcal{C})$ *has a discounting representation* $(w, D, \hat{D})$*. Then the individual is naive if and only if there exists* $\alpha \in [0, 1]$ *such that*

$$\hat{D}(t) = \alpha \frac{D(t+1)}{D(1)} + (1 - \alpha) D(t) \quad (\forall t \in \mathbb{N} \cup \{0\}), \tag{7}$$

and the individual is sophisticated if and only if $\alpha = 0$*. In addition, if the individual is strictly naive (i.e.,* $\alpha > 0$*), then*

1. *The discount function* D *exhibits diminishing impatience if and only if*

$$\frac{D(0)}{D(t)} > \frac{\hat{D}(0)}{\hat{D}(t)} \quad (\forall t \in \mathbb{N}).$$

2. The discount function D exhibits strong diminishing impatience if and only if

$$\frac{D(t)}{D(t+1)} > \frac{\hat{D}(t)}{\hat{D}(t+1)} \quad (\forall t \in \mathbb{N} \cup \{0\}).$$

The two equivalences under strict naiveté are surprising because they relate (strong) diminishing impatience of the actual temptation, as captured in D, with the intertemporal rate of substitution in the believed temptation, as captured in $\hat{D}$. The first claim says that for a strict naive individual, diminishing impatience is equivalent to beliefs being biased toward saving desirable consumption for a later date t rather than in the present period 0, as reflected in $\hat{D}(0)/\hat{D}(t) < D(0)/D(t)$. In other words, under-appreciating the temptation for immediate consumption versus later consumption is an inherent feature of naiveté with diminishing impatience. If beliefs are ever biased in the opposite direction (with projected undersaving) then the individual cannot exhibit diminishing impatience in her virtuous utility. Similarly, under-appreciation of the temptation to shift good consumption to immediately prior time periods is an inherent feature of strong diminishing impatience with naiveté. Note that the results do not suggest a relationship between the diminishing impatience of the actual temptation and the diminishing impatience of the believed temptation.

6 Welfare

Our behavioral definitions of naiveté provide a parsimonious language to conduct welfare analysis without relying on a particular representation. As its illustration, in this section we consider a setup that explores the welfare implications of policies that introduce new commitment devices to naive consumers. Suppose the government contemplates whether to provide an illiquid forced-savings device. This is equivalent to introducing an additional commitment device or menu x to the family of existing available menus; the new menu excludes immediate consumption beyond a certain level. A pervasive finding is that the take-up of new commitment devices is minimal under naiveté.[29] Beyond mitigating the effectiveness of new commitment devices, we find that new commitment devices can strictly decrease welfare when consumers are naive. In fact, the existence of such strictly deleterious commitment devices characterizes naiveté. Moreover, the marginal welfare effects of such interventions fail to be monotone in sophistication, except for the usage of complete commitments to a single outcome.

Formally, we consider families of menus to understand the effects of introducing additional commitment devices. For finite $X \subset \mathcal{K}(\Delta(C))$, let $x^*(X) = \{x \in X :$

[29] Several studies in this line are surveyed by Bryan, Karlan, and Nelson (2010).

$x \succsim y$ for all $y \in X\}$ denote the set of $\succsim$-maximal menus from the family X. Let $\mathfrak{C}(X) \in \mathcal{C}(x^*(X)) \equiv \{\mathcal{C}(x) : x \in x^*(X)\}$. That is, $\mathfrak{C}(X)$ is the final consumption from the family of menus X when the individual adheres to the following protocol: first, she selects a $\succsim$-maximal menu $x \in x^*(X)$, and second, she consumes $\mathcal{C}(x)$. This allows us to compare the final welfare from different families of commitment devices by comparing their induced final choices, that is, X is better for an individual than Y if $\{\mathfrak{C}(X)\} \succsim \{\mathfrak{C}(Y)\}$.[30] We focus on the deterministic case for simplicity, but the stochastic generalization is straightforward (except for Theorem 6 that makes explicit use of deterministic Strotz representations).

The next result makes the straightforward but important observation that adding additional commitment devices always makes sophisticated individuals better off. The converse result, that strictly naive individuals can always be made strictly worse off by introducing available commitments devices, requires that singleton menus are dense in the ex-ante preference as they are, for example, whenever a Strotz representation exists. The literature already observed many specific situations where providing flexibility to naive individuals makes them worse off. Our point is that this is a general phenomenon: the possibility of such welfare loss is a necessary consequence of strict naiveté. Recall that an individual is strictly naive if she is naive but not sophisticated.

Theorem 5. *If an individual is sophisticated, then* $\{\mathfrak{C}(X)\} \succsim \{\mathfrak{C}(Y)\}$ *whenever* $X \supset Y$. *If singleton menus are* $\succsim$*-dense and the individual is strictly naive, then there exist* $X \supset Y$ *with* $\{\mathfrak{C}(X)\} \prec \{\mathfrak{C}(Y)\}$.

The prior theorem is intuitive because the additional menu that leads to a less virtuous final selection is possibly a superset of an already available menu. Clearly, increasing flexibility for individuals who mistakenly believe they will virtuously exercise that flexibility can decrease their welfare. The next result is sharper, but requires the additional structure of the Strotz model. Under Strotz preferences, there always exists a subset of an existing menu that leaves the individual worse off when added to the family of commitments. That is, there exists a scenario where welfare is harmed by adding stronger commitments (rather than more flexibility) that exclude choices which are otherwise available in some existing forward plan in the status quo.

A natural question is whether the harmful effect of commitment devices is monotone with respect to the naiveté ordering. Excluding the extreme cases of sophistication and full naiveté, the second part of the result shows this is not the case. More specifically,

[30] We follow the commonly employed approach of using ex-ante commitment preferences over singletons as the welfare criterion over final consumption $\Delta(C)$. Another established benchmark is the Pareto welfare (partial) order based on improvements with respect to both ex-ante and ex-post preferences. Since Theorems 5 and 7 involve changing ex-ante utility u with possible reciprocal changes to ex-post utility v, they are no longer valid with respect to the Pareto welfare criterion. Theorem 6 involves losses to both ex-ante and ex-post utility, and therefore holds with respect to either welfare criterion.

for a generic pair of comparable individuals, there exists a new commitment device that leaves the more sophisticated individual strictly worse off while having no effect on the more naive individual.

Say an individual has a *preference for commitment* if there exist menus y and $x \subset y$ such that $x \succ y$. If an individual has a preference for commitment, then she is not fully naive in the sense of believing that her future tastes will be identical to her current commitment preference.

Theorem 6.

1. *Suppose $(\succsim, \mathcal{C})$ admits a Strotz representation $(u, v, \hat{v})$, where u and v are independent.*[31] *If the individual is strictly naive and has a preference for commitment, then there exist menus y and $x \subset y$ such that $\{\mathfrak{C}(\{x, y\})\} \prec \{\mathfrak{C}(\{y\})\}$.*

2. *Suppose $(\succsim_1, \mathcal{C}_1)$ and $(\succsim_2, \mathcal{C}_2)$ admit Strotz representations $(u, v_1, \hat{v}_1)$ and $(u, v_2, \hat{v}_2)$. Suppose individual 2 is strictly naive and has a preference for commitment, and that u_2 and v_2 are independent. If individual 1 is strictly more naive than individual 2,*[32] *then there exist menus y and $x \subset y$ such that $\{\mathfrak{C}_1(\{x, y\})\} \sim_1 \{\mathfrak{C}_1(\{y\})\}$ and $\{\mathfrak{C}_2(\{x, y\})\} \prec_2 \{\mathfrak{C}_2(\{y\})\}$.*

Example 1. For a concrete illustration of Theorem 6, consider an individual facing a three-period consumption-savings problem. In period 0, she initially chooses to invest money in a liquid savings account or in a retirement account. In period 1, she then decides whether to make a withdrawal from her savings. If she initially invested in the retirement account, then her early withdrawal results in a tax penalty. In period 2, she finally consumes the remaining balance of the savings or retirement account. For simplicity of exposition, we assume linear utility over static consumption and focus on deterministic consumption streams.

Suppose that $(\succsim, \mathcal{C})$ has a Strotz representation $(u, v, \hat{v})$ where

$$u(c_1, c_2) = c_1 + c_2, \quad v(c_1, c_2) = c_1 + \beta c_2, \quad \text{and} \quad \hat{v}(c_1, c_2) = c_1 + \hat{\beta} c_2.$$

As standard, assume $0 < \beta < \hat{\beta} \leq 1$. If the gross interest rate is $R > 1$ and the individual initially has unit wealth, then investing in the liquid savings account is equivalent to choosing the menu

$$y = \{(c_1, c_2) \in \mathbb{R}^2_+ : c_1 + c_2/R \leq 1\}.$$

[31] That is, it is not that case that $v \approx u$ or $v \approx -u$.

[32] That is, individual 1 is more naive than individual 2, but it is not the case that individual 2 is also more naive than individual 1. This restriction still permits a shared ex-ante or a shared ex-post Strotz representation, but not both.

The retirement account has a proportional early withdrawal penalty τc_1 associated with a withdrawal of c_1 in period 1, where $\tau \geq 0$. Thus investing in the retirement account is equivalent to choosing the menu

$$x^\tau = \{(c_1, c_2) \in \mathbb{R}^2_+ : (1+\tau)c_1 + c_2/R \leq 1\}.$$

Note that $x^\tau \subset y$, and that $x^\tau = y$ if $\tau = 0$. The actual choices from y and x^τ are

$$\mathcal{C}(y) = \begin{cases} (1,0) & \text{if } 1 > \beta R \\ (0,R) & \text{if } 1 \leq \beta R, \end{cases} \quad \text{and} \quad \mathcal{C}(x^\tau) = \begin{cases} (1/(1+\tau),0) & \text{if } 1 > (1+\tau)\beta R \\ (0,R) & \text{if } 1 \leq (1+\tau)\beta R. \end{cases}$$

Suppose $1 > \hat{\beta}R$ and $(1+\tau)\hat{\beta}R > 1 > (1+\tau)\beta R$. In this case, the individual correctly anticipates choosing $(1,0)$ from the menu y. However, she incorrectly anticipates choosing $(0,R)$ from x^τ, when in fact she will choose $(1/(1+\tau),0)$. She believes that the tax penalty associated with the retirement account is high enough to deter her from making early withdrawals in period 1, but in reality it is not. Since $u(0,R) > u(1,0)$, this incorrect belief will lead the individual to initially invest in the illiquid retirement account x^τ over the liquid savings account y in period 0. Therefore, the availability of the retirement account as a savings instrument is strictly detrimental, since

$$\{\mathfrak{C}(\{x^\tau, y\})\} = \{(1/(1+\tau),0)\} \prec \{(1,0)\} = \{\mathfrak{C}(\{y\})\}.$$

Finally, consider a more naive individual who is parameterized by $(\hat{\beta}^*, \beta^*)$ such that $\hat{\beta}^* R \geq 1$. This individual does not have a strict incentive to switch to the retirement account because she anticipates choosing $(0,R)$ even from y. Thus offering the commitment device to this individual has no harmful effect.[33] ■

Example 1 illustrates the potentially detrimental impact of offering additional "soft" commitments, that is, commitment devices that worsen tempting options by making them more expensive (e.g., retirement accounts with early withdrawal penalties). In a related field experiment by Giné, Karlan, and Zinman (2010), many participants who were offered commitment contracts for smoking cessation put money in a blocked account as a penalty for smoking, but failed to quit and lost their money.[34] In contrast, more naive

[33] It is possible that the more naive individual could select the retirement account in spite of her indifference between the two menus. However, modifying this example to include an arbitrarily small cost of setting up the retirement account will cause the more naive individual to strictly prefer the menu y without changing the less naive individual's strict preference for x^τ. Similarly, the conclusions in Theorem 6 in no way rely on how indifferences are broken.

[34] Relatedly, John (2016) offers commitment contracts to low-income households in the Philippines and lets them choose the default penalty. A majority of participants default on their contract, which suggests that they overestimated the effect of the penalty on their future behavior.

individuals would pass up these commitment devices under the false belief that they can quit smoking without any commitment, allowing them to avoid this penalty. Assuming free disposal of money, any "soft" commitment device that uses monetary penalties can be represented as a restriction on the set of feasible action/money pairs and hence takes the form of committing to a subset $x \subset y$ of the original option set, as in Theorem 6.[35]

The theme of Example 1 that soft commitment devices can prove harmful to partially naive individuals—and potentially more so as naiveté decreases—has been explored in other specific examples in the literature: Heidhues and Koszegi (2009) examine a setting where quasi-hyperbolic discounters can pay an up-front cost to impose a penalty on indulging future temptations. They show that welfare can fail to be monotonic in beliefs, with more accurate values of $\hat{\beta}$ sometimes leading to lower welfare. Spiegler (2011, Section 4.1.2) considers the choice of contract by a time-inconsistent consumer, with one contract delivering less favorable transfers but higher-powered incentives to choose the ex-ante preferred option. Similar to the main intuition in our savings example, he shows that decreases in naiveté may lead a consumer to choose the more expensive contract under the false belief that it will discipline future behavior, only to end up choosing the tempting option ex post at greater expense.[36] Theorem 6 shows the generality of these negative results: For *any* pair of partially naive individuals who are strictly ranked in terms of their naiveté, there exist a pair of commitment devices, one stronger than the other, such that making the stronger commitment device available is harmful only to the less naive individual. This result suggests caution when considering the distribution of welfare effects induced by policy changes affecting individuals with heterogeneous levels of naiveté.

While some forms of partial commitment can make naive individuals worse off, some classes of commitments can unambiguously improve welfare. In Example 1, increasing the early withdrawal penalty magnifies the strength of the commitment device x^τ. When τ is small and the commitment device is weak, the welfare effect of this commitment device is neutral or negative. Once τ exceeds some threshold, the exact value of which depends on the preference parameters, the commitment device becomes strong enough to discipline future behavior and deliver a positive welfare impact. Expanding on this

[35]In fact, a variation of Theorem 6 can be proven that shows the welfare loss for a partially naive individual can always take the form of a penalty that is foregone due to a failed attempt at self-control. Specifically, under the assumptions of part 1 of Theorem 6, there exists a menu y and lotteries $p \in y, q \notin y$ such that $\{p\} \succ \{q\}$, $y \cup \{q\} \setminus \{p\} \succ y$, $\mathcal{C}(y) = p$, and $\mathcal{C}(y \cup \{q\} \setminus \{p\}) = q$. For example, p is a harmful temptation and q is the same tempting good together with a penalty; the individual prefers trading p for q ex ante, but ultimately chooses the worsened option q ex post. The analogue of the comparative result in part 2 of Theorem 6 holds as well.

[36]In contrast, holding beliefs fixed, welfare is monotonically increasing with more virtuous actual behavior. Trivially, if two individuals share the same ex-ante preference ($u_1 \approx u_2$ and $\hat{v}_1 \approx \hat{v}_2$) and individual 2 is more virtuous ($v_2 \gg_u v_1$), then individual 2 is is better off in any *fixed* two-stage decision problem X than individual 1.

observation, the following result shows there is a class of commitment devices that will unambiguously improve welfare for any preference: complete commitments to a single outcome. In addition, the second part of the result shows that, in contrast to the case of partial commitments, the beneficial effect of offering complete commitments is monotone with respect to the naiveté ordering.

Theorem 7.

1. *Assume* $\mathfrak{C}(X \cup \{x\}) \notin x$ *implies* $\mathfrak{C}(X \cup \{x\}) = \mathfrak{C}(X)$.[37] *If an individual is naive, then* $\{\mathfrak{C}(X \cup \{p\})\} \succsim \{\mathfrak{C}(X)\}$ *for all lotteries* p.

2. *Suppose individuals 1 and 2 are naive and share the same commitment preference, and suppose 1 is more naive than 2. Then, for any menu* y *and lottery* p, $\{p\} \not\succ_1 y$ *and* $\{\mathfrak{C}_1(\{\{p\}, y\})\} \succ_1 \{\mathfrak{C}_1(\{y\})\}$ *imply* $\{\mathfrak{C}_2(\{\{p\}, y\})\} \succ_2 \{\mathfrak{C}_2(\{y\})\}$.

When menu choice is driven purely by temptation, adding extreme commitment devices is never harmful. Of course, such extreme commitment can be rejected if individuals have uncertain virtuous tastes that lead them to demand flexibility. Optimal design of commitment devices for naive individuals with some demand for flexibility remains an important open question.[38] In Section 7.3 we discuss the possibility of detecting naiveté in the presence of uncertain virtuous tastes.

7 Extensions

7.1 Comparative Naiveté with Non-Common Normative Preferences

In the previous sections, we have focused on comparing individuals who share a common normative ranking (commitment preference). In this section we relax this assumption and provide a generalization of our definition of comparative naiveté (in the deterministic case only). Recall that, in the deterministic case with common normative preferences, individual 1 is more naive than individual 2 if

$$x \succsim_2 \{p\} \succsim_2 \{\mathcal{C}_2(x)\} \implies x \succsim_1 \{p\} \succsim_1 \{\mathcal{C}_1(x)\}.$$

[37] This property simply implies that the tie-breaking procedure used in the selection function $\mathfrak{C}$ does not change when unchosen options are added. This avoids spurious welfare conclusions that are artifacts of the tie-breaking protocol.

[38] Amador, Werning, and Angeletos (2006) study a consumption-savings problem combining flexibility and temptation, but under the assumption of full sophistication.

Based on the same idea, the following definition also identifies gaps between commitment behavior and actual ex-post choices. Whenever the more sophisticated individual mistakenly prefers choosing later in the menu x to committing to her actual choice from x, which indicates a misprediction, the more naive individual displays the same behavior.

Definition 18. *Individual 1 is* more naive* *than individual 2 if, for all menus x,*

$$x \succ_2 \{\mathcal{C}_2(x)\} \implies x \succ_1 \{\mathcal{C}_1(x)\}.$$

Theorem 8. *Suppose $(\succsim_1, \mathcal{C}_1)$ and $(\succsim_2, \mathcal{C}_2)$ are strictly naive and have Strotz representations $(u_1, v_1, \hat{v}_1)$ and $(u_2, v_2, \hat{v}_2)$ such that individual 1 has a preference for commitment and $v_1 \not\approx \hat{v}_2$. Then individual 1 is* more naive* *than individual 2 if and only if there exist $\alpha, \alpha' \in [0, 1]$ such that*

$$v_2 \approx \alpha v_1 + (1-\alpha)\hat{v}_1 \quad \text{and} \quad \hat{v}_2 \approx \alpha' v_1 + (1-\alpha')\hat{v}_1.$$

Theorem 8 characterizes the implications of this definition in terms of deterministic Strotz representations, excluding special cases where tie-breaking can create spurious counter-examples.[39] The implication of comparative naiveté in this setting is that the discrepancy between $\hat{v}_1$ and v_1 is wider than that between $\hat{v}_2$ and v_2. Note that, in contrast to the common normative ranking case, the ranking $\hat{v}_1 \gg_{u_1} \hat{v}_2 \gg_{u_1} v_2 \gg_{u_1} v_1$ does not necessarily hold (the ordering of $\hat{v}_2$ and v_2 can be reversed). We leave an extension to the stochastic setting as an open question.

7.2 Costly Self-Control

So far we have not considered the possibility of an agent's costly effort to resist temptations. We now turn to analyzing the robustness of our results in the presence of such costly self-control. In particular, following Gul and Pesendorfer (2001), the individual's self-control cost of choosing alternative p from the menu x is $\max_{q \in x} v(q) - v(p)$, the difference between the temptation utility of the most tempting option and that of p. The individual maximizes her commitment utility u subject to these self-control costs, and therefore chooses the option that maximizes the compromise $u(p) + v(p)$ of commitment utility and temptation utility. The following definition permits uncertainty about the temptation utility, as in Stovall (2010), as well as possibility of incorrect beliefs about the distribution of temptation utilities.

Definition 19. *A* random Gul-Pesendorfer representation *of $(\succsim, \lambda)$ is a triple $(u, \mu, \hat{\mu})$ of a nontrivial expected-utility function u and nontrivial probability measures $\mu, \hat{\mu}$ over $\mathcal{V}$*

[39] Under $v_1 \approx \hat{v}_2$, the "if" direction might not hold. However, violations of the more naive* condition can occur only at particular menus where tie-breaking is needed for both individuals.

with finite-dimensional support such that the function $U : \mathcal{K}(\Delta(C)) \to \mathbb{R}$ *defined by*

$$U(x) = \int_{\mathcal{V}} \Big[\max_{p \in x}(u(p) + v(p)) - \max_{q \in x} v(q) \Big] d\hat{\mu}(v)$$

represents $\succsim$ *and*

$$\lambda^x(y) = \mu(p_x^{-1}(y))$$

for some measurable selection function $p_x : \mathcal{V} \to x$ *with* $p_x(v) \in B_{u+v}(x)$ *for all* $v \in V$.

Dekel and Lipman (2012) show that under a mild continuity assumption, the menu preference alone cannot distinguish the random Gul-Pesendorfer model and the random Strotz model. However, they also find that a random Gul-Pesendorfer representation of $\succsim$ implies different ex-post choice probabilities than those implied by a random Strotz representation of $\succsim$. Analyzing whether our results can be extended to deal with self-control preferences is therefore important since the identification of naiveté proposed in Section 3 relies on a particular model of commitment behavior.[40]

The following theorem states that if the individual is naive and if her behavior admits a random Gul-Pesendorfer representation, then the ex-ante beliefs derived from the representation are optimistic. More precisely, naiveté implies that any random Gul-Pesendorfer representation predicts ex-post choices that are more virtuous than the actual ones. The intuition is the following: Self-control costs increase the attractiveness of commitment since tempting options can be undesirable ex ante even if they are not chosen ex post. Thus the definition of absolute naiveté proposed in Section 3 serves as a conservative and robust test to reveal an individual's optimism even in the presence of costly self-control.

Theorem 9. *Suppose that* $(\succsim, \lambda)$ *has a random Gul-Pesendorfer representation* $(u, \mu, \hat{\mu})$*, and that the individual is naive. Then, for any* u*-upper set* $\mathcal{U}$*,*

$$\hat{\mu}(\{v \in \mathcal{V} \mid u + v \in \mathcal{U}\}) \geq \mu(\{v \in \mathcal{V} \mid u + v \in \mathcal{U}\}). \tag{8}$$

In addition, if the individual is strictly naive, Equation (8) *is satisfied with strict inequality for some* $\mathcal{U}$*.*

It is important to note that the converse of Theorem 9 fails. In particular, even if $\hat{\mu} = \mu$ and the individual has correct beliefs about her future behavior, the desire to avoid self-control costs may result in $\{m(\lambda^x)\} \succ x$ for some menus, in violation of both our behavioral definitions of naiveté and sophistication. Thus our behavioral definition

[40] Dekel and Lipman (2012) show that both representations can be distinguished based on ex-post choices under the assumption of sophistication. If the agent is naive, the models are obviously indistinguishable since the associated beliefs are revealed only by commitment preferences and not by ex-post choice frequencies.

of naiveté is sufficient but *not necessary* for overoptimistic beliefs when individuals have self-control preferences. In a companion paper Ahn, Iijima, and Sarver (2016), we explore alternative behavioral conditions that tightly characterize naiveté for deterministic self-control preferences. We also show that there is an impossibility result when randomness is permitted: It is impossible to construct a behavioral definition that tightly characterizes naiveté for the random Gul-Pesendorfer representation. Therefore, Theorem 9 is perhaps the best result that one can hope for when self-control costs are random.

7.3 Uncertainty in Normative Preferences

The random Strotz interpretation of commitment preferences relies on the assumption that normative preferences are certain ex ante. The elicitation of naiveté provided in Section 3 is therefore suited to situations where long-term preferences are known and where deviations are always undesirable (e.g., temptations, addictions, memory lapses). In some situations, however, the individual might expect future shocks to her normative preferences. In that case, her menu choices trade off commitment versus flexibility and the condition $x \succ \{m(\lambda^x)\}$ does not necessarily indicate unrealistic expectations: An individual who anticipates receiving some information about her normative ranking prior to selecting an option might rationally refuse to commit to her average choice.

Identifying the flexibility-loving part from the commitment-loving component of preferences in order to detect naive anticipations requires additional assumptions. For instance, Stovall (2014) assumes that the normative uncertainty (over u) is realized prior to the temptation uncertainty (over v), in which case the individual's beliefs can be identified from her preferences. In some contexts, the normative states are tied to objective contingencies that can be directly observed by the analyst (e.g., financial events, weather, health status). In both of these cases, the identification of naiveté can be performed as described in Section 3 conditional on each normative state.

In the general case, the sophistication hypothesis imposes some necessary properties on choice data. In particular, options can be relevant ex ante only if they are chosen with some probability ex post, an axiom that Ahn and Sarver (2013) call *consequentialism*: $x \sim \text{supp}(\lambda^x)$ for all menus x is a necessary condition for the existence of a sophisticated representation. In contrast, the condition $x \succ \text{supp}(\lambda^x)$ indicates that the individual overestimates the virtue of her choices inside x.

As an illustration, suppose that an individual is considering buying a membership that gives her free access to the gym. Let x denote the option set that includes any number of gym visits, and let $p \in x$ denote zero visits. Observing that she values the membership ex ante ($x \succ \{p\}$) but that she attends the gym with probability zero ex post ($\lambda^x(\{p\}) = 1$) is sufficient to conclude that she had unrealistic expectations regarding her

gym attendance.[41] Relatedly, suppose that the individual can self-impose a penalty for smoking, as described in Section 6. Her initial choice set is $\{p^1, p^2\}$ (smoking or not) but she can replace p^1 by a contract p^3 according to which smoking results in the payment of a penalty. Observing that she selects the contract ($\{p^3, p^2\} \succ \{p^1, p^2\}$) but continues smoking with probability one despite the penalty ($\lambda^{\{p^3,p^2\}}(\{p^3\}) = 1$) is sufficient to conclude that her menu choice was led by naive anticipations.[42] Note that in both of these examples, one needs to observe repeated observations for a single individual or a cross sectional distribution of a set of individuals who make the same ex-ante choices in order to determine the support of λ^x and infer naiveté.

[41] Note that all of the options in x are in fact pairs consisting of a number of visits together with the expense of the gym membership. Letting q denote zero visits without the paying for the membership, the choice to join the gym corresponds to the preference $x \succ \{q\}$. Since $\{q\} \succ \{p\}$ by dominance (the individual prefers not to pay the cost of the membership without going), we have $x \succ \{p\} = \text{supp}(\lambda^x)$.

[42] This argument implicitly assumes that the individual prefers having the option to quit to being forced to smoke ($\{p^1, p^2\} \succsim \{p^1\}$) and that the individual prefers not to pay the penalty all else equal ($\{p^1\} \succ \{p^3\}$). Thus $\{p^3, p^2\} \succ \{p^3\} = \text{supp}(\lambda^{\{p^3,p^2\}})$.

A A Comparative from Dekel and Lipman (2012)

In this section, we summarize a relevant result from Dekel and Lipman (2012) that will play a central role in our proofs of Theorems 2, 3, and 4. Recall the definition of comparative temptation aversion: Individual 2 is *more temptation averse* than individual 1 if, for all menus x and lotteries p,

$$\{p\} \succ_1 x \implies \{p\} \succ_2 x.$$

Theorem 10 (Dekel and Lipman (2012))**.** *Suppose $\succsim_1$ and $\succsim_2$ have random Strotz representations (u, μ_1) and (u, μ_2). Then $\succsim_2$ is more temptation averse than $\succsim_1$ if and only if $\mu_1 \gg_u \mu_2$.*

Dekel and Lipman (2012) consider only a finite prize space C in their paper. In the Supplemental Appendix, we prove that their result can be extended to any compact metric space C and any random Strotz representation (with finite-dimensional support) defined on that space.[43] This extension to compact spaces is not merely a technical exercise, as it is critical for many of the applications of our results, such as dynamic consumption problems where $C = [a, b]^{\mathbb{N}}$.

B Uncertain Intensity Random Strotz

In this section, we highlight a useful special case of the random Strotz representation where the uncertainty over future behavior is only over the magnitude of the future temptation, and not in its basic direction. For example, the individual may know that she will crave sweet snacks (but not salty snacks) ex post, but is uncertain of how strong her craving for sweets will be. This uncertain intensity Strotz representation encompasses the random quasi-hyperbolic discounting model studied in Section 5.1 where the individual is uncertain of the intensity of her present bias, and the corollaries presented below provide a bridge between our main theorems and the results in that section.

Recall that two expected-utility functions u and v are independent if they are nontrivial and it is not the case that $v \approx u$ or $v \approx -u$.

Definition 20. *An* uncertain intensity Strotz representation *of $(\succsim, \lambda)$ is a quadruple $(u, v, F, \hat{F})$ of two independent expected-utility functions u, v and two cumulative distribution functions $F, \hat{F}$ on $[0, 1]$ such that the function $U : \mathcal{K}(\Delta(C)) \to \mathbb{R}$ defined by*

$$U(x) = \int_0^1 \max_{p \in B_{\alpha u + (1-\alpha)v}(x)} u(p)\, d\hat{F}(\alpha)$$

[43] Definition 6 imposes the restriction that the measure μ in the random Strotz representation must have finite-dimensional support. It is an open question whether this comparative result can be extended to probably measures with arbitrary support. Our proof in the Supplemental Appendix shows that the "if" direction in Theorem 10 is true without the finite-dimensional support assumption. However, we view the exploration of additional generalizations of this results as a purely technical question. As we discussed in Section 3.2, we are not aware of any application of the random Strotz model that does not have finite-dimensional support.

is a utility representation of $\succsim$ *and, for all menus* x *and all measurable* $y \subset x$,

$$\lambda^x(y) = F(p_x^{-1}(y))$$

for some measurable selection function $p_x : [0,1] \to x$ *with* $p_x(\alpha) \in B_u(B_{\alpha u+(1-\alpha)v}(x))$ *for all* $\alpha \in [0,1]$.

For the case of an uncertain intensity Strotz representation, the direction of the temptation is known to be v, but the magnitude of that temptation relative to the virtuous utility u is uncertain. A naive individual underestimates the influence of v, and this bias is reflected in her belief $\hat{F}$ over the intensities in $[0,1]$ putting more likelihood on larger weighting of u (hence lower weighting of v) than F.

Corollary 10. *Suppose* $(\succsim, \lambda)$ *has a uncertain intensity Strotz representation* $(u, v, F, \hat{F})$. *Then the individual is naive if and only if* $\hat{F} \geq_{FOSD} F$ *(and is sophisticated if and only if* $\hat{F} = F$*).*

Corollary 11. *Suppose* $(\succsim_1, \lambda_1)$ *and* $(\succsim_2, \lambda_2)$ *are naive and have uncertain intensity Strotz representations* $(u, v, F_1, \hat{F}_1)$ *and* $(u, v, F_2, \hat{F}_2)$.

1. *Individual 1 is more naive than individual 2 if and only if*

$$\hat{F}_1 \geq_{FOSD} \hat{F}_2 \geq_{FOSD} F_2 \geq_{FOSD} F_1.$$

2. *Individual 1 is weakly more naive than individual 2 if and only if*

$$F_1(\alpha) - \hat{F}_1(\alpha) \geq F_2(\alpha) - \hat{F}_2(\alpha), \quad \forall \alpha \in [0,1].$$

C Proofs

C.1 Proof of Theorem 2

Suppose the random choice rule λ has a random Strotz representation (u, μ). Consider the hypothetical sophisticated ex-ante preference $\succsim^*$ that is also be represented by (u, μ). The following lemma shows how this hypothetical preference can be determined from λ and u.

Lemma 2. *Suppose* λ *has a random Strotz representation* (u, μ). *Then for any menu* x,

$$u(m(\lambda^x)) = \int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\mu(v).$$

In particular, if we define a binary relation $\succsim^*$ *on* $\mathcal{K}(\Delta(C))$ *by*

$$x \succsim^* y \iff u(m(\lambda^x)) \geq u(m(\lambda^y)),$$

then (u, μ) *is a random Strotz representation for* $\succsim^*$.

Proof. If (u, μ) represents λ then by definition there exists, for all menus x, a measurable selection function $p_x : \mathcal{V} \to x$ with $p_x(v) \in B_u(B_v(x))$ such that

$$\lambda^x(y) = \mu(p_x^{-1}(y))$$

for all measurable $y \subset x$. Thus λ^x is the distribution on x induced by the random variable p_x defined on the measure space $(\mathcal{V}, \mu)$. Therefore, the standard change of variables formula together with the linearity and continuity of u imply

$$\begin{aligned}\int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\mu(v) &= \int_{\mathcal{V}} u(p_x(v))\, d\mu(v) \\ &= \int_x u(p)\, d\lambda^x(p) = u\bigg(\int_x p\, d\lambda^x(p)\bigg) = u(m(\lambda^x)),\end{aligned}$$

as desired. ■

Turning now to the proof of Theorem 2, fix a random Strotz representation $(u, \mu, \hat{\mu})$ for $(\succsim, \lambda)$, and define $\succsim^*$ as in Lemma 2. To establish sufficiency, suppose the individual is naive. Then for all menus x and lotteries p,

$$\begin{aligned}\{p\} \succ x &\implies \{p\} \succ \{m(\lambda^x)\} \qquad \text{(naiveté)} \\ &\implies u\big(m\big(\lambda^{\{p\}}\big)\big) = u(p) > u(m(\lambda^x)) \\ &\implies \{p\} \succ^* x.\end{aligned}$$

Thus $\succsim^*$ is more temptation averse than $\succsim$. Since (u, μ) represents $\succsim^*$ by Lemma 2, Theorem 10 implies that $\hat{\mu} \gg_u \mu$. If the individual is sophisticated, then a similar argument shows that the converse also holds: $\succsim$ is also more temptation averse than $\succsim^*$ (in particular, $\succsim = \succsim^*$) and hence $\mu \gg_u \hat{\mu}$ also holds, i.e., $\hat{\mu} \approx \mu$.

To establish necessity, suppose $\hat{\mu} \gg_u \mu$. By Theorem 10, $\succsim^*$ is more temptation averse than $\succsim$. By contrapositive, this is equivalent to the condition

$$x \succsim^* \{p\} \implies x \succsim \{p\}.$$

Note that for any menu x, if we take $p = m(\lambda^x)$ then

$$u(m(\lambda^x)) = u(p) = u\big(m\big(\lambda^{\{p\}}\big)\big)$$

and hence $x \sim^* \{p\} = \{m(\lambda^x)\}$. Since $\succsim^*$ is more temptation averse than $\succsim$, this implies $x \succsim \{m(\lambda^x)\}$. Thus the individual is naive. If we also have $\mu \gg_u \hat{\mu}$ then another application of Theorem 10 implies the condition above can be strengthened to $x \succsim^* \{p\} \iff x \succsim \{p\}$. In this case, $x \sim^* \{m(\lambda^x)\}$ implies $x \sim \{m(\lambda^x)\}$ and hence the individual is sophisticated.

C.2 Proof of Theorem 3

The following lemma decomposes our definition of more naive into two more basic conditions. The comparative of being more temptation averse is defined in the main text. The comparative of being more virtuous is defined for the first time in this lemma. Intuitively, individual 2 is more virtuous than individual 1 if she makes "better" choices (as measured by her commitment preference) from every menu than individual 1.

Lemma 3. *Suppose* $(\succsim_1, \lambda_1)$ *and* $(\succsim_2, \lambda_2)$ *are naive, and suppose* $\succsim_1$ *and* $\succsim_2$ *share the same commitment preference, i.e.,* $\{p\} \succsim_1 \{q\} \iff \{p\} \succsim_2 \{q\}$ *for all lotteries* $p, q \in \Delta(C)$*. Then individual 1 is more naive than individual 2 if and only if both of the following hold:*

1. *Individual 2 is more temptation averse than individual 1:* $\{p\} \succ_1 x \implies \{p\} \succ_2 x$.
2. *Individual 2 is more virtuous than individual 1:* $\{p\} \succ_2 \{m(\lambda_2^x)\} \implies \{p\} \succ_1 \{m(\lambda_1^x)\}$.

Proof. More naive implies (1): Fix any menu x and lottery p such that $\{p\} \succ_1 x$. Since individual 1 is more naive than 2, we cannot have $x \succsim_2 \{p\} \succsim_2 \{m(\lambda_2^x)\}$. Thus either $\{p\} \succ_2 x$ or $\{m(\lambda_2^x)\} \succ_2 \{p\}$. To rule out the second possibility, note that since individual 2 is naive, we must have $x \succsim_2 \{m(\lambda_2^x)\}$. Since individual 1 is more naive than 2, this implies $x \succsim_1 \{m(\lambda_2^x)\} \succsim_1 \{m(\lambda_1^x)\}$. Therefore, $\{p\} \succ_1 \{m(\lambda_2^x)\}$, and hence $\{p\} \succ_2 \{m(\lambda_2^x)\}$ since individuals 1 and 2 have the same commitment preference. Thus the only possibility is $\{p\} \succ_2 x$, as desired.

More naive implies (2): Fix any menu x and lottery p such that $\{p\} \succ_2 \{m(\lambda_2^x)\}$. Since individual 2 in naive, $x \succsim_2 \{m(\lambda_2^x)\}$. Since individual 1 is more naive than 2, this implies $x \succsim_1 \{m(\lambda_2^x)\} \succsim_1 \{m(\lambda_1^x)\}$. Individuals 1 and 2 share the same commitment preference, and therefore $\{p\} \succ_1 \{m(\lambda_2^x)\} \succsim_1 \{m(\lambda_1^x)\}$, as desired.

(1) and (2) together imply more naive: If individual 2 is more virtuous than individual 1, then we must have $\{m(\lambda_2^x)\} \succsim_2 \{m(\lambda_1^x)\}$. Otherwise, taking $p = m(\lambda_1^x)$ in condition (2) would lead to a contradiction. Therefore, since the individuals share the same commitment preference, $\{p\} \succsim_2 \{m(\lambda_2^x)\} \implies \{p\} \succsim_1 \{m(\lambda_1^x)\}$ for any lottery p. Combining this with the contrapositive of condition (1), it follows directly that individual 1 is more naive than individual 2. ■

We are now ready to prove Theorem 3. By Theorem 2, individual 2 is naive if and only if $\hat{\mu}_2 \gg_u \mu_2$. Also by Lemma 3, individual 1 is more naive than individual 2 if and only if 2 is both more temptation averse and more virtuous than 1. By Theorem 10, individual 2 is more temptation averse than individual 1 if and only if $\hat{\mu}_1 \gg_u \hat{\mu}_2$. The proof is therefore completed if we can show that individual 2 is more virtuous than individual 1 if and only if $\mu_2 \gg_u \mu_1$. To see that this is true, define $\succsim_1^*$ and $\succsim_2^*$ as in Lemma 2 for λ_1 and λ_2, respectively. Then (u, μ_1) and (u, μ_2) represent $\succsim_1^*$ and $\succsim_2^*$. Note that for all menus x and lotteries p,

$$\{p\} \succ_i \{m(\lambda_i^x)\} \iff u(p) > u(m(\lambda_i^x)) \iff \{p\} \succ_i^* x, \quad i = 1, 2.$$

Therefore, individual 2 is more virtuous than individual 1 if and only if $\succsim_1^*$ is more temptation averse than $\succsim_2^*$. By Theorem 10, this is true if and only if $\mu_2 \gg_u \mu_1$.

C.3 Proof of Lemma 1

Fix any x and p, q such that

$$\frac{1}{2}x + \frac{1}{2}\{p\} \succsim_2 \frac{1}{2}\{m(\lambda_2^x)\} + \frac{1}{2}\{q\}.$$

Take a lottery r such that $x \sim_2 \{r\}$. By independence,

$$\frac{1}{2}\{r\} + \frac{1}{2}\{p\} \sim_2 \frac{1}{2}x + \frac{1}{2}\{p\} \succsim_2 \frac{1}{2}\{m(\lambda_2^x)\} + \frac{1}{2}\{q\}.$$

Note also that $x \succsim_2 \{r\} \succsim_2 \{m(\lambda_2^x)\}$ since 2 is naive, and hence Definition 10 implies $x \succsim_1 \{r\} \succsim_1 \{m(\lambda_2^x)\} \succsim_1 \{m(\lambda_1^x)\}$. By independence, and since $\succsim_1$ and $\succsim_2$ share the same commitment preference,

$$\frac{1}{2}x + \frac{1}{2}\{p\} \succsim_1 \frac{1}{2}\{r\} + \frac{1}{2}\{p\} \succsim_1 \frac{1}{2}\{m(\lambda_2^x)\} + \frac{1}{2}\{q\} \succsim_1 \frac{1}{2}\{m(\lambda_1^x)\} + \frac{1}{2}\{q\},$$

as desired.

C.4 Proof of Theorem 4

(1) ⇔ (3): Let U_i denote the value function from the representation $(u, \hat{\mu}_i)$ for the ex-ante preference $\succsim_i$ for $i = 1, 2$. Also, recall from Lemma 2 that if λ_i has random Strotz representation (u, μ_i), then

$$u(m(\lambda_i^x)) = \int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\mu_i(v).$$

Thus for any menu x and lotteries p, q,

$$\begin{aligned} \frac{1}{2}x + \frac{1}{2}\{p\} \succsim_i \frac{1}{2}\{m(\lambda_i^x)\} + \frac{1}{2}\{q\} &\iff \frac{1}{2}U_i(x) + \frac{1}{2}u(p) \geq \frac{1}{2}u(m(\lambda_i^x)) + \frac{1}{2}u(q) \\ &\iff OV_i(x) \geq u(q) - u(p). \end{aligned}$$

It follows directly from this observation that individual 1 is weakly more naive than individual 2 if and only if $OV_1(x) \geq OV_2(x)$ for all x.

(2) ⇔ (3): Fix any lotteries p, q with $\{q\} \succ_i \{p\}$ for $i = 1, 2$. For each menu x, define

$$\begin{aligned} A_i^x &\equiv \big\{\alpha \in [0,1] : (1-\alpha)x + \alpha\{p\} \succsim_i (1-\alpha)\{m(\lambda_i^x)\} + \alpha\{q\}\big\} \\ &= \big\{\alpha \in [0,1] : (1-\alpha)OV_i(x) \geq \alpha(u(q) - u(p))\big\}. \end{aligned}$$

By definition, $P_i(x; p, q) = \sup A_i^x$. Note that A_i^x is a closed interval. Moreover, since both individuals are naive, we have $x \succsim_i m(\lambda_i^x)$ and therefore $0 \in A_i^x$. Also, $1 \notin A_i^x$ since $\{q\} \succ_i \{p\}$. This implies

$$\alpha = P_i(x; p, q) \iff (1-\alpha)OV_i(x) = \alpha(u(q) - u(p)).$$

Therefore, $OV_1(x) \geq OV_2(x)$ if and only if $P_1(x; p, q) \geq P_2(x; p, q)$.

(3) ⇔ (4): For any menu x,

$$\begin{aligned} &OV_1(x) \geq OV_2(x) \\ &\quad\iff \int_{\mathcal{V}} \max_{p\in B_v(x)} u(p)\, d\hat{\mu}_1(v) - \int_{\mathcal{V}} \max_{p\in B_v(x)} u(p)\, d\mu_1(v) \\ &\qquad\qquad \geq \int_{\mathcal{V}} \max_{p\in B_v(x)} u(p)\, d\hat{\mu}_2(v) - \int_{\mathcal{V}} \max_{p\in B_v(x)} u(p)\, d\mu_2(v) \\ &\quad\iff \int_{\mathcal{V}} \max_{p\in B_v(x)} u(p)\, d\Big(\frac{1}{2}\hat{\mu}_1 + \frac{1}{2}\mu_2\Big)(v) \geq \int_{\mathcal{V}} \max_{p\in B_v(x)} u(p)\, d\Big(\frac{1}{2}\hat{\mu}_2 + \frac{1}{2}\mu_1\Big)(v). \end{aligned}$$

If this is true of all menus x, then the (hypothetical) preference represented by the random Strotz representation $(u, \frac{1}{2}\hat{\mu}_2 + \frac{1}{2}\mu_1)$ is more temptation averse than the preference represented by $(u, \frac{1}{2}\hat{\mu}_1 + \frac{1}{2}\mu_2)$. Thus by Theorem 10, $OV_1(x) \geq OV_2(x)$ for all x if and only if $\frac{1}{2}\hat{\mu}_1 + \frac{1}{2}\mu_2 \gg_u \frac{1}{2}\hat{\mu}_2 + \frac{1}{2}\mu_1$ or, equivalently,

$$\frac{1}{2}\hat{\mu}_1(\mathcal{U}) + \frac{1}{2}\mu_2(\mathcal{U}) \geq \frac{1}{2}\hat{\mu}_2(\mathcal{U}) + \frac{1}{2}\mu_1(\mathcal{U})$$

for every u-upper set $\mathcal{U}$. Rearranging terms, this is precisely condition (4).

C.5 Proof of Corollaries 6 and 7

A maximally present-biased preference only values immediate consumption in period 1 and ignores all subsequent consumption, which is equivalent to the extreme case where $\beta = 0$: $v_0(c) = w(c_1)$. Any convex combination of the virtuous utility u and maximally present-biased v_0 can be rewritten as the following familiar formula:

$$\beta u(c) + (1-\beta)v_0(c) = w(c_1) + \beta \sum_{t=2}^{\infty} \delta^{t-1} w(c_t).$$

Therefore, uncertainty about the present-bias parameter β simply parameterizes uncertainty about the intensity of u relative to v_0, and β is the relative weighting of exponential discounting versus extreme impatience. Thus an RQH representation $(w, F, \hat{F}, \delta)$ can equivalently be expressed as an uncertain intensity Strotz representation $(u, v_0, F, \hat{F})$. With this observation, the results follow directly from Corollaries 10 and 11 in Appendix B.

C.6 Proof of Corollary 9

By Theorem 1, the individual is naive if and only if $\hat{v} \approx \alpha u + (1-\alpha)v$ for some $\alpha \in [0,1]$, where u, v, and $\hat{v}$ satisfy Equations (4), (5), and (6). This is equivalent to the condition

$$\hat{D}(t-1) = a\big[\alpha D(t) + (1-\alpha)D(t-1)\big], \quad \forall t \in \mathbb{N},$$

for some $a > 0$. Since $D(0) = \hat{D}(0) = 1$, we have $1/a = \alpha D(1) + (1-\alpha)$ and therefore

$$\begin{aligned}\hat{D}(t-1) &= \frac{\alpha D(t) + (1-\alpha)D(t-1)}{\alpha D(1) + (1-\alpha)} \\ &= \alpha' \frac{D(t)}{D(1)} + (1-\alpha')D(t-1),\end{aligned}$$

where

$$\alpha' = \frac{\alpha D(1)}{\alpha D(1) + (1-\alpha)} \in [0,1].$$

This establishes that the individual is naive if and only if Equation (7) is satisfied.

To prove claim 1, note that by Equation (7),

$$\hat{D}(t) > D(t) \iff \frac{D(t+1)}{D(1)} > D(t) = \frac{D(t)}{D(0)}.$$

The latter holds for all $t \in \mathbb{N}$ if and only if D exhibits diminishing impatience.

To prove claim 2, note first that

$$\begin{aligned}\frac{\hat{D}(t)}{\hat{D}(t+1)} &= \frac{\alpha \frac{D(t+1)}{D(1)} + (1-\alpha)D(t)}{\alpha \frac{D(t+2)}{D(1)} + (1-\alpha)D(t+1)} \\ &= \frac{\alpha \frac{D(t+1)}{D(t)} + (1-\alpha)D(1)}{\alpha \frac{D(t+2)}{D(t+1)} + (1-\alpha)D(1)} \cdot \frac{D(t)}{D(t+1)}.\end{aligned}$$

Therefore,

$$\frac{\hat{D}(t)}{\hat{D}(t+1)} < \frac{D(t)}{D(t+1)} \iff \frac{D(t+1)}{D(t)} < \frac{D(t+2)}{D(t+1)}.$$

The latter holds for all $t \in \mathbb{N} \cup \{0\}$ if and only if D exhibits strong diminishing impatience.

C.7 Proof of Theorem 5

By the standard revealed-preference argument, $X \supset Y$ implies $x \succsim y$ for any $x \in x^*(X)$ and $y \in x^*(Y)$. Under sophistication, $\{\mathcal{C}(x)\} \sim x \succsim y \sim \{\mathcal{C}(y)\}$. But $\mathfrak{C}(X) = \mathcal{C}(x)$ for some $x \in x^*(X)$ and $\mathfrak{C}(Y) = \mathcal{C}(y)$ for some $y \in x^*(Y)$, so in particular $\{\mathfrak{C}(X)\} \succsim \{\mathfrak{C}(Y)\}$.

Now assume the individual is strictly naive: There exists a menu x with $x \succ \{\mathcal{C}(x)\}$. By $\succsim$-denseness of the singletons, there exists some lottery p such that $x \succ \{p\} \succ \{\mathcal{C}(x)\}$. Let $X = \{x, \{p\}\}$ and $Y = \{\{p\}\}$. Then $\mathfrak{C}(Y) = p$ and $\mathfrak{C}(X) = \mathcal{C}(x)$, so $\{\mathfrak{C}(Y)\} \succ \{\mathfrak{C}(X)\}$.

C.8 Proof of Theorem 6

First part: Under the assumptions of the theorem, it can be shown that there exist lotteries p^1, p^2, p^3 such that[44]

$$\begin{aligned} &u(p^1) > u(p^2) > u(p^3) \\ &\hat{v}(p^2) > \hat{v}(p^1) > \hat{v}(p^3) \\ &v(p^2) > v(p^3) > v(p^1). \end{aligned}$$

Let $y = \{p^1, p^2, p^3\}$ and $x = \{p^1, p^3\}$. The rankings of the lotteries according to u and $\hat{v}$ imply that $x \sim \{p^1\} \succ \{p^2\} \sim y$. The ranking according to v implies that $\mathcal{C}(x) = p^3$ and $\mathcal{C}(y) = p^2$. Therefore, $\{\mathfrak{C}(\{x, y\})\} = \{p^3\} \prec \{p^2\} = \{\mathfrak{C}(\{y\})\}$.

Second part: By Corollary 2, $\hat{v}_1 \gg_u \hat{v}_2 \gg_u v_2 \gg_u v_1$. There are two cases to consider, depending on whether $v_1 \approx v_2$ or not.

Case 1—$v_1 \approx v_2$: Let $v \equiv v_2 \approx v_1$. Since individual 1 is strictly more naive than individual 2, in this case we must have $\hat{v}_1 \gg_u \hat{v}_2$, but not $\hat{v}_1 \approx \hat{v}_2$. Therefore, it can be shown that there exist lotteries p^1, p^2, p^3, p^4 such that[45]

$$\begin{aligned} &u(p^1) > u(p^2) > u(p^3) > u(p^4) \\ &\hat{v}_1(p^1) > \hat{v}_1(p^2) > \hat{v}_1(p^3) > \hat{v}_1(p^4) \\ &\hat{v}_2(p^3) > \hat{v}_2(p^1), \hat{v}_2(p^2) > \hat{v}_2(p^4) \\ &v(p^3) > v(p^4) > v(p^1), v(p^2). \end{aligned}$$

Let $y = \{p^1, p^2, p^3, p^4\}$ and $x = \{p^2, p^4\}$. The rankings of the lotteries according to u and $\hat{v}_1, \hat{v}_2$ imply that $y \sim_1 \{p^1\} \succ_1 \{p^2\} \sim_1 x$ and $y \sim_2 \{p^3\} \prec_2 \{p^2\} \sim_2 x$. The ranking according to v implies that $\mathcal{C}_i(y) = p^3$ and $\mathcal{C}_i(x) = p^4$ for $i = 1, 2$. Therefore, $\{\mathfrak{C}_1(\{x, y\})\} = \{p^3\} = \{\mathfrak{C}_1(\{y\})\}$ and $\{\mathfrak{C}_2(\{x, y\})\} = \{p^4\} \prec_2 \{p^3\} = \{\mathfrak{C}_2(\{y\})\}$.

Case 2—v_1 is not an affine transformation of v_2: Under these assumptions, it can be shown

[44]Proof: By Theorem 1, $\hat{v} \gg_u v$. Since it is not the case that $v \approx -u$, this implies $\hat{v} \approx \alpha u + (1-\alpha)v$. Note that $\alpha > 0$ since the individual is strictly naive, and $\alpha < 1$ since $\succsim$ has preference for commitment. Hence, it is not the case that $\hat{v} \approx u$, so there exist lotteries p, q such that $\hat{v}(p) = \hat{v}(q)$ and $u(p) > u(q)$. Since $\hat{v} \approx \alpha u + (1-\alpha)v$ for $\alpha \in (0, 1)$, this also implies that $v(p) < v(q)$. Since it is not the case that $v \approx -u$, there exist lotteries r, s such that $u(r) > u(s)$ and $v(r) > v(s)$, which also implies $\hat{v}(r) > \hat{v}(s)$. It is easy to show that the lotteries $p^1 = (1-\varepsilon)p + \varepsilon[(1/2)s + (1/2)r]$, $p^2 = (1-\varepsilon)q + \varepsilon r$, $p^3 = (1-\varepsilon)q + \varepsilon s$ have the desired properties for $\varepsilon > 0$ sufficiently small.

[45]The arguments needed to prove this claim are similar to those in footnote 44 and are omitted.

that there exist lotteries p^1, p^2, p^3 such that

$$\begin{aligned}
&u(p^1) > u(p^2) > u(p^3)\\
&\hat{v}_2(p^2) > \hat{v}_2(p^1) > \hat{v}_2(p^3)\\
&v_2(p^2) > v_2(p^3) > v_2(p^1)\\
&v_1(p^3) > v_1(p^2) > v_1(p^1).
\end{aligned}$$

The ranking of these lotteries according to $\hat{v}_1$ is not important for the result, although it is true that the above rankings and $\hat{v}_1 \gg_u \hat{v}_2$ imply $\hat{v}_1(p^1), \hat{v}_1(p^2) > \hat{v}_1(p^3)$. Let $y = \{p^1, p^2, p^3\}$ and $x = \{p^1, p^3\}$. The ranking according to v_1 implies $\mathcal{C}_1(y) = \mathcal{C}_1(x) = p^3$, so $\{\mathfrak{C}_1(\{x, y\})\} = \{p^3\} = \{\mathfrak{C}_1(\{y\})\}$. The rankings according to u and $\hat{v}_2$ imply that $y \sim_2 \{p^2\} \prec_2 \{p^1\} \sim_2 x$. The ranking according to v_2 implies that $\mathcal{C}_2(y) = p^2$ and $\mathcal{C}_2(x) = p^3$. Thus $\{\mathfrak{C}_2(\{x, y\})\} = \{p^3\} \prec_2 \{p^2\} = \{\mathfrak{C}_2(\{y\})\}$.

C.9 Proof of Theorem 7

First part: By the assumed properties of $\mathfrak{C}$, either $\mathfrak{C}(X \cup \{p\}) = \mathfrak{C}(X)$, in which case the results holds trivially, or $\mathfrak{C}(X \cup \{p\}) = p$. In the latter case, we must have $\{p\} \succsim x$ for all $x \in X$. Since $\mathfrak{C}(X) = \mathcal{C}(y)$ for some $y \in x^*(X)$, naiveté then implies $\{p\} \succsim y \succsim \{\mathcal{C}(y)\}$, and hence $\{\mathfrak{C}(X \cup \{p\})\} \succsim \{\mathfrak{C}(X)\}$.

Second part: The conditions $\{p\} \not\succ_1 y$ and $\{\mathfrak{C}_1(\{\{p\}, y\})\} \succ_1 \{\mathfrak{C}_1(\{y\})\}$ imply $\{p\} \succ_1 y$. Since individual 1 is more naive than 2, by Lemma 3 individual 2 is more temptation averse than 1 and therefore $\{p\} \succ_1 y$ implies $\{p\} \succ_2 y$. In addition, since 2 is naive, $y \succsim_2 \{\mathcal{C}_2(y)\}$. Thus $\{p\} \succ_2 y \succsim_2 \{\mathcal{C}_2(y)\}$ which in turn implies $\{\mathfrak{C}_2(\{\{p\}, y\})\} \succ_2 \{\mathfrak{C}_2(\{y\})\}$.

C.10 Proof of Theorem 8

We first show the "only if" part, which is divided into two steps. The claim in Step 2 implies the desired form by setting $\alpha = \frac{\phi}{\phi+\hat{\phi}}$ and $\alpha' = \frac{\psi}{\psi+\hat{\psi}}$.

Step 1: $v_2 \approx \phi v_1 + \hat{\phi}\hat{v}_1$ and $\hat{v}_2 \approx \psi v_1 + \hat{\psi}\hat{v}_1$ hold for some numbers $\phi, \hat{\phi}, \psi, \hat{\psi}$.

We first claim that v_2 is affine equivalent to a linear combination of $\hat{v}_2$ and u_1. If not, since $v_2 \not\approx \hat{v}_2$ (by the strict naiveté of 2), we can find p, q such that $u_1(p) = u_1(q)$, $\hat{v}_2(p) > \hat{v}_2(q)$, and $v_2(p) < v_2(q)$ by the standard argument. This implies $u_2(p) > u_2(q)$ by the naiveté of 2, and thus $\{p, q\} \succ_2 \{\mathcal{C}_2(\{p, q\})\}$. But $\{p, q\} \sim_1 \{\mathcal{C}_1(\{p, q\})\}$, which is a contradiction.

We next claim that $\hat{v}_2$ is affine equivalent to a linear combination of v_2 and v_1. If not, as in the above paragraph, we can find p, q such that $v_1(p) = v_1(q)$, $\hat{v}_2(p) > \hat{v}_2(q)$, and $v_2(p) < v_2(q)$ by the standard argument. Then $\mathcal{C}_1(\{p, q\}) \in \operatorname{argmax}_{r \in \{p,q\}} u_1(r)$ and thus $\{p, q\} \sim_1 \{\mathcal{C}_1(\{p, q\})\}$. Since $\{p, q\} \succ_2 \{\mathcal{C}_2(\{p, q\})\}$, it leads to a contradiction.

Since u_1 is affine equivalent to a linear combination of v_1 and $\hat{v}_1$ (as 1 is naive), the above two claims imply the desired formulas.

Step 2: $\phi, \hat{\phi}, \psi, \hat{\psi} \geq 0$ such that $\phi + \hat{\phi}, \psi + \hat{\psi} > 0$.

If either $\phi < 0$ or $\hat{\phi} < 0$, because of $v_2 \not\approx \hat{v}_2$ (as 2 is strictly naive), we can find p, q from the interior of $\Delta(C)$ such that $\hat{v}_1(p) > \hat{v}_1(q)$, $v_1(p) > v_1(q)$, and $v_2(p) < v_2(q)$. First consider the case $\hat{v}_2(p) \geq \hat{v}_2(q)$. This implies $u_2(p) \geq u_2(q)$ since 2 is strictly naive. Then we can find q' close to q such that strict inequalities $\hat{v}_1(p) > \hat{v}_1(q')$, $v_1(p) > v_1(q')$, $u_2(p) > u_2(q')$, $\hat{v}_2(p) > \hat{v}_2(q')$, and $v_2(p) < v_2(q')$ hold. We have $\{p, q'\} \sim_1 \{\mathcal{C}_1(\{p, q'\})\}$ and $\{p, q'\} \succ_2 \{\mathcal{C}_2(\{p, q'\})\}$, a contradiction. Next consider the case $\hat{v}_2(p) < \hat{v}_2(q)$. Then, because of $v_2 \not\approx \hat{v}_2$ (as 2 is strictly naive), we can find q' close to q such that $\hat{v}_2(q) > \hat{v}_2(q')$ and $v_2(q) < v_2(q')$ hold. Take such q' to be sufficiently close to q such that strict inequalities $\hat{v}_1(p) > \hat{v}_1(q')$, $v_1(p) > v_1(q')$, and $\hat{v}_2(p) < \hat{v}_2(q')$ hold. Then we have $\{p, q, q'\} \sim_1 \{\mathcal{C}_1(\{p, q, q'\})\}$ and $\{p, q, q'\} \succ_2 \{\mathcal{C}_2(\{p, q, q'\})\}$, a contradiction.

If either $\psi < 0$ or $\hat{\psi} < 0$, then there exist p, q such that $\hat{v}_1(p) > \hat{v}_1(q)$, $v_1(p) > v_1(q)$, and $\hat{v}_2(p) < \hat{v}_2(q)$. Then, depending on $v_2(p) \geq v_2(q)$ or $v_2(p) < v_2(q)$, an analogous argument as in the previous paragraph leads to a contradiction to the assumption that 1 is more naive* than 2.

We next show the "if" part. Take any x such that $x \succ_2 \{\mathcal{C}_2(x)\}$. Denote $p_i = \mathcal{C}_i(x) \in B_{u_i}(B_{v_i}(x))$ and $\hat{p}_i \in B_{u_i}(B_{\hat{v}_i}(x))$ for each $i = 1, 2$. We will show $u_1(\hat{p}_1) > u_1(p_1)$, which ensures $x \succ_1 \{\mathcal{C}_1(x)\}$. There are two cases to consider.

Case 1: $\hat{v}_2 \gg_{u_1} v_2$.

First, we have $\hat{v}_1(\hat{p}_1) \geq \hat{v}_1(\hat{p}_2)$ and $\hat{v}_2(\hat{p}_1) \leq \hat{v}_2(\hat{p}_2)$. If $\hat{v}_1(\hat{p}_1) = \hat{v}_1(\hat{p}_2)$ then $u_1(\hat{p}_1) \geq u_1(\hat{p}_2)$, since $p_1 \in B_{u_1}(B_{\hat{v}_1}(x))$. If $\hat{v}_1(\hat{p}_1) > \hat{v}_1(\hat{p}_2)$ then $u_1(\hat{p}_1) > u_1(\hat{p}_2)$ holds, since $\hat{v}_1 \gg_{u_1} \hat{v}_2$.

Second, we have $\hat{v}_2(\hat{p}_2) \geq \hat{v}_2(p_2)$ and $v_2(\hat{p}_2) \leq v_2(p_2)$. Because at least one of them is strict (otherwise $x \succ_2 \{\mathcal{C}_2(x)\}$ would not hold) and $\hat{v}_2 \approx \alpha u_1 + (1 - \alpha) v_2$ with $\alpha \in [0, 1)$, $u_1(\hat{p}_2) > u_1(p_2)$ follows.

Third, we have $v_2(p_2) \geq v_2(p_1)$ and $v_1(p_2) \leq v_1(p_1)$. If $v_2(p_2) = v_2(p_1)$ then $u_2(p_2) \geq u_2(p_1)$ since $p_2 \in B_{u_2}(B_{v_2}(x))$. This implies $u_1(p_2) \geq u_1(p_1)$ because either $u_2 \approx \alpha u_1 + (1 - \alpha) v_2$ or $u_1 \approx \alpha u_2 + (1 - \alpha) v_2$ with $\alpha \in (0, 1]$. If $v_2(p_2) > v_2(p_1)$ then $u_1(p_2) > u_1(p_1)$ since $v_2 \gg_{u_1} v_1$.

Case 2: $v_2 \gg_{u_1} \hat{v}_2$.

This case is almost analogous to the previous case. First, we have $\hat{v}_1(\hat{p}_1) \geq \hat{v}_1(p_2)$ and $v_2(\hat{p}_1) \leq v_2(p_2)$. If $\hat{v}_1(\hat{p}_1) = \hat{v}_1(p_2)$, then $u_1(\hat{p}_1) \geq u_1(p_2)$, since $p_1 \in B_{u_1}(B_{\hat{v}_1}(x))$. If $\hat{v}_1(\hat{p}_1) > \hat{v}_1(p_2)$ then $u_1(\hat{p}_1) > u_1(p_2)$ holds, since $\hat{v}_1 \gg_{u_1} v_2$.

Second, we have $v_2(p_2) \geq v_2(\hat{p}_2)$ and $\hat{v}_2(p_2) \leq \hat{v}_2(\hat{p}_2)$. Because at least one of them is strict and $v_2 \approx \alpha u_1 + (1 - \alpha) \hat{v}_2$ with $\alpha \in [0, 1)$, $u_1(p_2) > u_1(\hat{p}_2)$ follows.

Third, we have $\hat{v}_2(\hat{p}_2) \geq \hat{v}_2(p_1)$ and $v_1(\hat{p}_2) \leq v_1(p_1)$. Because $\hat{v}_2 \approx \alpha u_1 + (1 - \alpha) v_1$ with $\alpha \in (0, 1]$ (by assumption $\hat{v}_2 \not\approx v_1$), $u_1(\hat{p}_2) \geq u_1(p_1)$ follows.

C.11 Proof of Theorem 9

Define the function $\sigma : \mathcal{V} \to \mathcal{V}$ by $\sigma(v) = u + v$, and define the measures $\hat{\nu}$ and ν on $\mathcal{V}$ by $\hat{\nu}(E) = \hat{\mu}(\sigma^{-1}(E))$ and $\nu(E) = \mu(\sigma^{-1}(E))$ for any measurable set E. Observe that for any menu x,

$$\begin{aligned}
\int_{\mathcal{V}} \min_{p \in B_v(x)} u(p) d\hat{\nu}(v) &= \int_{\mathcal{V}} \min_{p \in B_{u+v}(x)} u(p) d\hat{\mu}(v) \qquad \text{(change of variables)} \\
&\geq \int_{\mathcal{V}} \Big[\max_{p \in x} (u(p) + v(p)) - \max_{q \in x} v(q) \Big] d\hat{\mu}(v) \\
&= U(x) \\
&\geq u(m(\lambda^x)) \qquad \text{(naiveté)} \\
&= \int_{\mathcal{V}} u(p_x(v)) d\mu(v) \\
&\geq \int_{\mathcal{V}} \min_{p \in B_{u+v}(x)} u(p) d\mu(v) \\
&= \int_{\mathcal{V}} \min_{p \in B_v(x)} u(p) d\nu(v). \qquad \text{(change of variables)}
\end{aligned} \tag{9}$$

Thus,

$$\int_{\mathcal{V}} \min_{p \in B_v(x)} u(p) d\hat{\nu}(v) \geq \int_{\mathcal{V}} \min_{p \in B_v(x)} u(p) d\nu(v),$$

which we rewrite as

$$\int_{\mathcal{V}} \max_{p \in B_v(x)} [-u(p)] d\hat{\nu}(v) \leq \int_{\mathcal{V}} \max_{p \in B_v(x)} [-u(p)] d\nu(v). \tag{10}$$

Consider the binary relations $\succsim^{\hat{\nu}}$ and $\succsim^{\nu}$ defined by their Random Strotz representations $(-u, \hat{\nu})$ and $(-u, \nu)$, respectively. Equation (10) shows that $\succsim^{\hat{\nu}}$ is more temptation-averse than $\succsim^{\nu}$. Theorem 10 applies since $\hat{\nu}$ and ν have finite-dimensional supports, and implies that $\nu \gg_{-u} \hat{\nu}$.

Consider a u-upper set $\mathcal{U}$, and $v \in \mathcal{V} \setminus \mathcal{U}$, $v' \in \mathcal{V}$ such that $v' \gg_{-u} v$. It is easy to show that this latter condition is equivalent to $v \gg_u v'$. Suppose that $v' \in \mathcal{U}$. Since $\mathcal{U}$ is a u-upper set, the condition $v \gg_u v'$ implies $v \in \mathcal{U}$, which is a contradiction. Hence, $v' \in \mathcal{V} \setminus \mathcal{U}$ for any v' such that $v' \gg_{-u} v$. This shows that $\mathcal{V} \setminus \mathcal{U}$ is a $(-u)$-upper set, and therefore $\nu(\mathcal{V} \setminus \mathcal{U}) \geq \hat{\nu}(\mathcal{V} \setminus \mathcal{U})$, or equivalently $\hat{\nu}(\mathcal{U}) \geq \nu(\mathcal{U})$.

We therefore have

$$\hat{\mu}(\{v \in \mathcal{V} \mid u + v \in \mathcal{U}\}) = \hat{\nu}(\mathcal{U}) \geq \nu(\mathcal{U}) = \mu(\{v \in \mathcal{V} \mid u + v \in \mathcal{U}\}). \tag{11}$$

To complete the proof, we show that Equation (11) is strict for some $\mathcal{U}$ if the individual is strictly naive. Suppose, by contradiction, that Equation (11) is satisfied as an equality for all u-upper sets. The arguments above imply that $\hat{\nu}(\mathcal{U}) = \nu(\mathcal{U})$ for any $(-u)$-upper set $\mathcal{U}$, i.e.,

by Theorem 10 that $\succsim^{\hat{\nu}}$ is more temptation-averse than $\succsim^{\nu}$ and vice versa. This implies that Equation (10) is satisfied as an equality for all x, and therefore the system in Equation (9) only contains equalities. In particular, $U(x) = u(m(\lambda^x))$ for all x, i.e., the individual is sophisticated.

C.12 Proof of Corollary 10

Lemma 4. *Suppose u and v are independent expected-utility functions, and define a function $g : [0,1] \to \mathcal{V}$ by $g(\alpha) = \alpha u + (1-\alpha)v$.*

1. *Take any cumulative distribution functions F and $\hat{F}$ on $[0,1]$, and define probability measures μ and $\hat{\mu}$ on $\mathcal{V}$ by $\mu \equiv F \circ g^{-1}$ and $\hat{\mu} \equiv \hat{F} \circ g^{-1}$.[46] If $(u, v, F, \hat{F})$ is an uncertain intensity Strotz representation of a preference $(\succsim, \lambda)$, then $(u, \mu, \hat{\mu})$ is a random Strotz representation of $(\succsim, \lambda)$.*

2. *Take any cumulative distribution functions F_1 and F_2 on $[0,1]$, and define probability measures μ_1 and μ_2 on $\mathcal{V}$ by $\mu_i \equiv F_i \circ g^{-1}$. Then $\mu_1 \gg_u \mu_2$ if and only if $F_1 \geq_{FOSD} F_2$.*

Proof. (1): Note that by assumption $\succsim$ is represented by

$$U(x) = \int_0^1 \max\{u(p) : p \in B_{g(\alpha)}(x)\}\, d\hat{F}(\alpha).$$

By the standard change of variables formula, this implies

$$\begin{aligned} U(x) &= \int_{\mathcal{V}} \max\{u(p) : p \in B_{\tilde{v}}(x)\}\, d(\hat{F} \circ g^{-1})(\tilde{v}) \\ &= \int_{\mathcal{V}} \max\{u(p) : p \in B_{\tilde{v}}(x)\}\, d\hat{\mu}(\tilde{v}), \end{aligned}$$

and hence $(u, \hat{\mu})$ is a random Strotz representation of $\succsim$.

Note also that by assumption there exists, for each menu x, a measurable selection function $p_x : [0,1] \to x$ with $p_x(\alpha) \in B_u(B_{g(\alpha)}(x))$ for all $\alpha \in [0,1]$ such that

$$\lambda^x(y) = F(p_x^{-1}(y))$$

for all measurable $y \subset x$. Take any measurable selection function $\tilde{p}_x : \mathcal{V} \to x$ with $\tilde{p}_x(\tilde{v}) \in B_u(B_{\tilde{v}}(x))$ for all $\tilde{v} \in \mathcal{V}$ that also satisfies $p_x(\alpha) = \tilde{p}_x(g(\alpha))$ for all $\alpha \in [0,1]$.[47] Therefore, for

[46] We are abusing notation slightly and using F to also denote the probability measure on $[0,1]$ that has F as its distribution function. That is, for any measurable set $A \subset [0,1]$, we write $F(A)$ to denote $\int_A dF(\alpha)$. Thus $\mu(E) = \int_{\{\alpha' : g(\alpha') \in E\}} dF(\alpha)$ for any measurable $E \subset \mathcal{V}$.

[47] To see that such a selection function $\tilde{p}_x$ exists, fix any measurable selection function $\hat{p}_x : \mathcal{V} \to x$ with $\hat{p}_x(\tilde{v}) \in B_u(B_{\tilde{v}}(x))$ for all $\tilde{v} \in \mathcal{V}$. Let $\bar{\mathcal{V}} = g([0,1]) \subset \mathcal{V}$. When the codomain of g is restricted to $\bar{\mathcal{V}}$, i.e., $g : [0,1] \to \bar{\mathcal{V}}$, this function is a bijection. Now define $\tilde{p}_x(\tilde{v}) = p_x(g^{-1}(\tilde{v}))$ for $\tilde{v} \in \bar{\mathcal{V}}$ and $\tilde{p}_x(\tilde{v}) = \hat{p}_x(\tilde{v})$ for $\tilde{v} \notin \bar{\mathcal{V}}$.

any measurable $y \subset x$,

$$\lambda^x(y) = F(g^{-1}(\tilde{p}_x^{-1}(y))) = \mu(\tilde{p}_x^{-1}(y)),$$

and hence (u, μ) is a random Strotz representation of λ.

(2): Suppose $\mu_i \equiv F_i \circ g^{-1}$ for $i = 1, 2$ and $\mu_1 \gg_u \mu_2$. Fix any $\alpha \in [0, 1]$, and let $\mathcal{U} = \{v' \in \mathcal{V} : v' \gg_u \alpha u + (1-\alpha)v\}$. By construction, $\mathcal{U}$ is a u-upper set, so $\mu_1(\mathcal{U}) \geq \mu_2(\mathcal{U})$. In addition, $g^{-1}(\mathcal{U}) = [\alpha, 1]$. Therefore,

$$F_1([\alpha, 1]) = \mu_1(\mathcal{U}) \geq \mu_2(\mathcal{U}) = F_2([\alpha, 1]).$$

Since this is true for all $\alpha \in [0, 1]$, $F_1 \geq_{FOSD} F_2$.

Conversely, suppose $F_1 \geq_{FOSD} F_2$. Fix any u-upper set $\mathcal{U}$. Note that for any $0 \leq \alpha \leq \alpha' \leq 1$, we have $g(\alpha') \gg_u g(\alpha)$ and hence

$$g(\alpha) \in \mathcal{U} \implies g(\alpha') \in \mathcal{U}.$$

This implies that the set $g^{-1}(\mathcal{U})$ is an interval from some $\alpha^* \in [0, 1]$ to 1.[48] Therefore,

$$\mu_1(\mathcal{U}) = F_1(g^{-1}(\mathcal{U})) \geq F_2(g^{-1}(\mathcal{U})) = \mu_2(\mathcal{U}).$$

Since this is true for all u-upper sets, $\mu_1 \gg_u \mu_2$. ■

Turning now to the proof of Corollary 10, suppose $(\succsim, \lambda)$ has an uncertain intensity Strotz representation $(u, v, F, \hat{F})$. Define g as in Lemma 4 for u and v, define measures $\mu \equiv F \circ g^{-1}$ and $\hat{\mu} \equiv \hat{F} \circ g^{-1}$ on $\mathcal{V}$. By part 1 of Lemma 4, $(u, \mu, \hat{\mu})$ is a random Strotz representation for $(\succsim, \lambda)$. Therefore, by Theorem 2 together with part 2 of Lemma 4, the individual is naive if and only if $\hat{F} \geq_{FOSD} F$ (and is sophisticated if and only if $\hat{F} = F$).

C.13 Proof of Corollary 11

Suppose $(\succsim_1, \lambda_1)$ and $(\succsim_2, \lambda_2)$ are naive and have uncertain intensity Strotz representations $(u, v, F_1, \hat{F}_1)$ and $(u, v, F_2, \hat{F}_2)$. Define g as in Lemma 4 for u and v, define measures $\mu_i \equiv F_i \circ g^{-1}$ and $\hat{\mu}_i \equiv \hat{F}_i \circ g^{-1}$ on $\mathcal{V}$. By part 1 of Lemma 4, $(u, \mu_i, \hat{\mu}_i)$ is a random Strotz representation for $(\succsim_i, \lambda_i)$ for $i = 1, 2$. The result follows from applications of Theorems 3 and 4, respectively, together with part 2 of Lemma 4.

[48] That is, it is equal to either $(\alpha^*, 1]$ or $[\alpha^*, 1]$, where $\alpha^* = \inf g^{-1}(\mathcal{U})$.

References

Ahn, D. S. (2007): "Ambiguity Without a State Space," *Review of Economic Studies*, 75, 3–28.

Ahn, D. S., R. Iijima, and T. Sarver (2016): "Naiveté about Temptation and Self-Control: Foundations for Naive Quasi-Hyperbolic Discounting," Working paper.

Ahn, D. S., and T. Sarver (2013): "Preference for Flexibility and Random Choice," *Econometrica*, 81, 341–361.

Ahn, D. S., and T. Sarver (2015): "Comparative Measures of Naiveté," Economic Research Initiatives at Duke (ERID) Working Paper No. 186. Available at SSRN: http://ssrn.com/abstract=2600209

Ali, S. N. (2011): "Learning Self-Control," *Quarterly Journal of Economics*, 126, 857–893.

Aliprantis, C., and K. Border (2006): *Infinite Dimensional Analysis*, 3rd edition. Berlin, Germany: Springer-Verlag.

Amador, M., I. Werning, and G.-M. Angeletos. "Commitment vs. Flexibility," *Econometrica*, 74, 365–396.

Augenblick, N., M. Niederle, and C. Sprenger (2015): "Working Over Time: Dynamic Inconsistency in Real Effort Tasks," *Quarlery Journal of Economics*, 130, 1067–1115.

Augenblick, N., and M. Rabin (2015): "An Experiment on Time Preference and Misprediction in Unpleasant Tasks," Working paper, Haas School of Business and Harvard University.

Bryan, G., D. Karlan, and S. Nelson (2010): "Commitment Devices," *Annual Review of Economics*, 2, 671–698.

DellaVigna, S. (2009): "Psychology and Economics: Evidence from the Field," *Journal of Economic Literature*, 47, 315–372.

DellaVigna, S., and U. Malmendier (2006): "Paying Not to Go to the Gym," *American Economic Review*, 96, 694–719.

Dekel, E., and B. L. Lipman (2010): "Costly Self-Control and Random Self-Indulgence," Working paper.

Dekel, E., and B. L. Lipman (2012): "Costly Self-Control and Random Self-Indulgence," *Econometrica*, 80, 1271–1302.

Dekel, E., B. L. Lipman, and A. Rustichini (2009): "Temptation-Driven Preferences," *Review of Economic Studies*, 76, 9371–971

Duflo, E., M. Kremer, and J. Robinson (2011): "Nudging Farmers to Use Fertilizer: Evidence from Kenya," *American Economic Review*, 101, 2350–2390.

Eliaz, K., and R. Spiegler (2006): "Contracting with Diversely Naive Agents," *Review of Economic Studies*, 72, 689–714.

Epstein, L. G. (1999): "A Definition of Uncertainty Aversion," *Review of Economic Studies*, 66, 579-608.

Freeman, D. (2016): "Revealing Sophistication and Naïveté from Procrastination," Working paper, Simon Fraser University.

Giné, X., D. Karlan, and J. Zinman (2010): "Put your Money where your Butt is: a Commitment Contract for Smoking Cessation," *American Economic Journal: Applied Economics*, 2, 213-235.

Ghirardato, P., and M. Marinacci (2002): "Ambiguity Made Precise: A Comparative Foundation," *Journal of Economic Theory*, 102, 251-289.

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

Gul, F., and W. Pesendorfer (2004): "Self-Control and the Theory of Consumption," *Econometrica*, 72, 119–158.

Gul, F., and W. Pesendorfer (2005): "The Revealed Preference Theory of Changing Tastes," *Review of Economic Studies*, 72, 429–448.

Gul, F., and W. Pesendorfer (2006): "Random Expected Utility," *Econometrica*, 74, 121–146.

Heidhues, P., and B. Koszegi (2009): "Futile Attempts at Self-Control," *Journal of the European Economic Association*, 7, 423–434.

Heidhues, P., and B. Koszegi (2010): "Exploiting Naivete about Self-Control in the Credit Market," *American Economic Review*, 100, 2279–2303.

John, A. (2016): "When Commitment Fails — Evidence from a Field Experiment," Working paper.

Kaur, S., M. Kremer, and S. Mullainathn (2015): "Self-Control at Work," *Journal of Political Economy*, 123, 1227–1277.

Kopylov, I. (2012): "Perfectionism and Choice," *Econometrica*, 80, 1819–1943.

Koszegi, B. (2014): "Behavioral Contract Theory," *Journal of Economic Literature*, 52, 1075–1118.

Kreps, D., and E. Porteus (1978): "Temporal Resolution of Uncertainty and Dynamic Choice Theory," *Econometrica*, 46, 185–200.

Krusell, P., B. Kuruşçu, and A. Smith (2010): "Temptation and Taxation," *Econometrica*, 78, 2063–2084.

Le Yaouanq, Y. (2015): "Anticipating Preference Reversal," Toulouse School of Economics Working Paper No. TSE-585.

Lipman, B. L., and W. Pesendorfer (2013): "Temptation," in Acemoglu, Arellano, and Dekel, eds., *Advances in Economics and Econometrics: Tenth World Congress*, Volume 1, Cambridge University Press.

Madrian, B. C., and D. F. Shea (2001): "The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior," *Quarterly Journal of Economics*, 116, 1149–1187.

Noor, J. (2007): "Commitment and Self-Control," *Journal of Economic Theory*, 135, 1-34.

Noor, J. (2011): "Temptation and Revealed Preference," *Econometrica*, 79, 601–644.

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

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

Peleg, M., and M. E. Yaari (1973): "On the Existence of a Consistent Course of Action when Tastes are Changing," *Review of Economic Studies*, 40, 391–401.

Prelec, D. (2004): "Decreasing Impatience: A Criterion for Non-stationary Time Preference and 'Hyperbolic' Discounting," *Scandinavian Journal of Economics*, 106, 511–532.

Sarver, T. (2008): "Anticipating Regret: Why Fewer Options May Be Better," *Econometrica*, 76, 263–305.

Shui, H., and L. M. Ausubel (2005): "Time Inconsistency in the Credit Card Market," Working paper, University of Maryland.

Spiegler, R. (2011): *Bounded Rationality and Industrial Organization*. New York, NY: Oxford University Press.

Stovall, J. (2010): "Multiple Temptations," *Econometrica*, 78, 349-376.

Stovall, J. (2014): "Temptation with uncertain normative preferences," Working paper.

Supplementary Appendix for

Behavioral Characterizations of Naiveté for Time-Inconsistent Preferences

S.1 Pessimism about Self-Control

While our main focus is on naiveté in the traditional sense of underestimation of future temptations, simple variations of our definitions can be used to model an individual who overestimates her future temptations and is therefore overly cautious. In this section, we summarize the implications of such pessimistic violations of sophistication. Formal results are stated for the case the deterministic Strotz representation for simplicity, but the analogous results for random choice are also true.

Definition S.1. *An individual is* pessimistic *if* $\{\mathcal{C}(x)\} \succsim x$ *for all menus* x*.*

An individual who is pessimistic has an actual temptation utility than is more aligned with her normative utility than her anticipated temptation utility.

Theorem S.1. *Suppose* $(\succsim, \mathcal{C})$ *has a Strotz representation* $(u, v, \hat{v})$*. Then the individual is pessimistic if and only if* $v \gg_u \hat{v}$*.*

The proof of this result is similar to that of Theorem 1 and is omitted.

Definition S.2. *Individual 1 is* more pessimistic *than individual 2 if, for all menus* x *and lotteries* p*,*

$$\{\mathcal{C}_2(x)\} \succsim_2 \{p\} \succsim_2 x \implies \{\mathcal{C}_1(x)\} \succsim_1 \{p\} \succsim_1 x.$$

In contrast to the case of individual 1 being more naive than individual 2, now individual 1 accepts more unnecessary (detrimental) commitments than individual 2. This comparative corresponds to a reversal of the ordering of temptation utilities obtained in Corollary 2.

Theorem S.2. *Suppose* $(\succsim_1, \mathcal{C}_1)$ *and* $(\succsim_2, \mathcal{C}_2)$ *are pessimistic and have Strotz representations* $(u, v_1, \hat{v}_1)$ *and* $(u, v_2, \hat{v}_2)$*. Then individual* 1 *is more pessimistic than individual* 2 *if and only if*

$$v_1 \gg_u v_2 \gg_u \hat{v}_2 \gg_u \hat{v}_1.$$

S.2 Proof of Theorem 10

S.2.1 Sufficiency: more temptation averse $\Longrightarrow$ less u-aligned

The following is the relevant result from Dekel and Lipman (2012), which they proved for the case of finite C.

Theorem S.3 (Dekel and Lipman (2012))**.** *Suppose C has finite cardinality. Suppose $\succsim_1$ and $\succsim_2$ have random Strotz representations (u, μ_1) and (u, μ_2). Then $\succsim_2$ is more temptation averse than $\succsim_1$ if and only if $\mu_1 \gg_u \mu_2$.*

Proof. Theorem 4 in Dekel and Lipman (2012) establishes the equivalence of $\succsim_2$ being more temptation averse than $\succsim_1$ and another condition on the representations that they refer to as conditional dominance. However, they also establish that $\mu_1 \gg_u \mu_2$ as an intermediate step in their proof.[49] The equivalence asserted in Theorem 10 is also stated explicitly in Theorem 4 of their working paper, Dekel and Lipman (2010).[50] ■

To prove the sufficiency part of Theorem 10, we now show that the sufficiency direction in Theorem S.3 can be extended to any compact and metrizable space C and any random Strotz representations (u, μ_1) and (u, μ_2) defined on that space, subject to our restriction that each μ_i has finite-dimensional support. Our approach is to show that the relationship between μ_1 and μ_2, specifically $\mu_1 \gg_u \mu_2$, can be inferred from looking at the restriction of the representations and preferences to a carefully chosen finite consumption space $C^* \subset C$.

The following preliminary result will be useful in the sequel. Recall that $\mathcal{V}$ denotes the set of all continuous functions $v : C \to \mathbb{R}$, i.e., the set of all expected-utility functions.

Lemma S.1. *Suppose the set $\{v_1, \ldots, v_n\} \subset \mathcal{V}$ is linearly independent. Then there exists a finite subset $C^* \subset C$ such that the set $\{v_1^*, \ldots, v_n^*\}$ is linearly independent, where $v_i^* = v_i|_{C^*}$ is the restriction of the function v_i to C^*.*

[49] To show that $\succsim_2$ being more temptation averse that $\succsim_1$ implies $\mu_1 \gg_u \mu_2$, the relevant results in Dekel and Lipman (2012) are the following: Lemma 3 shows that a partial order vC_uv' used in their paper is equivalent to our order $v \gg_u v'$ (ignoring their normalization of utility functions). Lemmas 4, 5, and 6 and the arguments on page 1296 show that for any set W that is closed under C_u (is a u-upper set in our terminology), $\mu_1(W) \geq \mu_2(W)$.

[50] Dekel and Lipman (2010) impose a normalization on the set of utility functions used in their result. However, by the uniqueness properties of the random Strotz representation established in Theorem 3 of Dekel and Lipman (2012), the probability of any u-upper set is the same for any random Strotz representation of the same preference. Therefore, their normalization of utilities is inconsequential for the result.

Proof. Suppose to the contrary that for every finite $B \subset C$, the collection $\{v_1|_B, \ldots, v_n|_B\}$ is linearly dependent. Then for any finite $B \subset C$, the set $A_B \subset \mathbb{R}^n$ defined by

$$A_B = \{\alpha \in \mathbb{R}^n : \|\alpha\| = 1 \text{ and } \alpha_1 v_1(c) + \cdots + \alpha_n v_n(c) = 0 \ \forall c \in B\}$$

is nonempty. Note that A_B is also a closed subset of the unit ball in $\mathbb{R}^n$, which is itself compact because n is finite. Let $\mathcal{B}$ denote the set of all nonempty finite subsets of C. For any $B_1, \ldots, B_k \in \mathcal{B}$, we have

$$A_{B_1} \cap \cdots \cap A_{B_k} = A_{B_1 \cup \cdots \cup B_k} \neq \emptyset,$$

since $B_1 \cup \cdots \cup B_k$ is finite and hence also in $\mathcal{B}$. Thus the collection $\{A_B\}_{B \in \mathcal{B}}$ has the finite intersection property. Since these sets are closed subsets of a compact set, this implies $\bigcap_{B \in \mathcal{B}} A_B \neq \emptyset$. However, since

$$\bigcap_{B \in \mathcal{B}} A_B = \{\alpha \in \mathbb{R}^n : \|\alpha\| = 1 \text{ and } \alpha_1 v_1(c) + \cdots + \alpha_n v_n(c) = 0 \ \forall c \in C\},$$

this implies the set $\{v_1, \ldots, v_n\}$ is linearly dependent, a contradiction. ■

Since μ_1 and μ_2 have finite-dimensional support, there exists a finite set of expected-utility functions $\{v_1, \ldots, v_n\} \subset \mathcal{V}$ such that $\text{supp}(\mu_i) \subset \text{span}(\{v_1, \ldots, v_n\})$ for $i = 1, 2$. Consider the set of function $\{u, \mathbf{1}, v_1, \ldots, v_n\}$, where $\mathbf{1}$ denotes the constant function with $\mathbf{1}(c) = 1$ for all $c \in C$. Without loss of generality, assume that this set of functions is linearly independent. Otherwise, we can sequentially remove the functions v_i until we obtain a linearly independent set.[51] To simplify notation in what follows, let $\mathcal{V}_s \equiv \text{span}(\{u, \mathbf{1}, v_1, \ldots, v_n\}) \subset \mathcal{V}$. Thus $\mu_1(\mathcal{V}_s) = \mu_2(\mathcal{V}_s) = 1$.

Take C^* as in Lemma S.1 for the set $\{u, \mathbf{1}, v_1, \ldots, v_n\}$. Let $\mathcal{V}^*$ denote the set of all continuous real-valued functions on C^* and let $\mathcal{V}_s^* \equiv \text{span}(\{u^*, \mathbf{1}^*, v_1^*, \ldots, v_n^*\}) \subset \mathcal{V}^*$, where $u^* = u|_{C^*}$, $\mathbf{1}^* = \mathbf{1}|_{C^*}$, and $v_i^* = v_i|_{C^*}$. Note that each of the functions $u^*, v_1^*, \ldots, v_n^*$ must be nontrivial (i.e., not constant) since function $\mathbf{1}^*$ together with these functions forms a linearly independent set.

Lemma S.2. *Define a function $g : \mathcal{V}_s \to \mathcal{V}_s^*$ by $g(v) = v|_{C^*}$, and define a measure μ_i^* on $\mathcal{V}^*$ by $\mu_i^*(E) = \mu_i(g^{-1}(E))$ for any measurable set $E \subset \mathcal{V}^*$ for $i = 1, 2$.*[52]

[51] Note that the set $\{u, \mathbf{1}\}$ must be linearly independent since u assumed to be nontrivial (i.e., not constant). Moreover, if $\text{span}\{u, \mathbf{1}\} = \text{span}\{u, \mathbf{1}, v_1, \ldots, v_n\}$, then the support of the measures in the random Strotz representations (u, μ_i) must assign all probability to the set of affine transformations of u. In this case, the representations reduce to time-consistent expected-utility maximization, and we have $\mu_1 \approx \mu_2$. Except in this trivial case, the linearly independent set of expected-utility functions whose span contains the support of μ_i must contain u, $\mathbf{1}$, and at least some of the v_i functions.

[52] In the definition of μ_i^*, we are implicitly treating g as a function from $\mathcal{V}_s$ into $\mathcal{V}^*$. We could equivalently define μ_i^* by $\mu_i^*(E) = \mu_i(g^{-1}(E \cap \mathcal{V}_s^*))$.

1. *The function g is a homeomorphism. That is, g is bijection and both g and its inverse function g^{-1} are continuous.*

2. *For any measurable set $E \subset \mathcal{V}$, $\mu_i(E) = \mu_i^*(g(E \cap \mathcal{V}_s))$.*

3. *For any proper u-upper set $\mathcal{U}$ in $\mathcal{V}$ (i.e., $\mathcal{U} \subsetneq \mathcal{V}$), the set $\mathcal{U}^* = g(\mathcal{U} \cap \mathcal{V}_s)$ is a u^*-upper set in $\mathcal{V}^*$.*

4. *Let $\succsim_i^*$ denote the restriction of $\succsim_i$ to sets of lotteries with support in C^*, which we can identify with the set $\mathcal{K}(\Delta(C^*))$. Then (u^*, μ_i^*) is a random Strotz representation for $\succsim_i^*$ for $i = 1, 2$.*

Proof. (1): This is a standard application of the fundamental theorem linear algebra for finite-dimensional vector spaces. Note that g is a linear function from the linear space $\mathcal{V}_s$ with basis vectors $\{u, \mathbf{1}, v_1, \ldots, v_n\}$ to the linear space $\mathcal{V}_s^*$ with basis vectors $\{u^*, \mathbf{1}^*, v_1^*, \ldots, v_n^*\}$. Since g maps each basis vector for $\mathcal{V}_s$ to the corresponding basis vector for $\mathcal{V}_s^*$ and the number of basis vectors is the same for each space, g is a bijection. Since any linear function between finite-dimensional spaces is continuous, both g and g^{-1} are continuous.[53]

(2): Fix any measurable set $E \subset \mathcal{V}$. Then

$$\mu_i(E) = \mu_i(E \cap \mathcal{V}_s) = \mu_i(g^{-1}(g(E \cap \mathcal{V}_s))) = \mu_i^*(g(E \cap \mathcal{V}_s)),$$

where the first equality follows from $\mu_i(\mathcal{V}_s) = 1$, the second follows from $g^{-1}(g(E \cap \mathcal{V}_s)) = E \cap \mathcal{V}_s$ (which holds because g is a bijection), and the third follows from the definition of μ_i^*.

(3): First observe that for any $v, v' \in \mathcal{V}_s$,

$$\begin{aligned} v \approx v' &\iff v = av' + b\mathbf{1} \text{ for some } a > 0, b \in \mathbb{R} \\ &\iff g(v) = ag(v') + b\mathbf{1} \text{ for some } a > 0, b \in \mathbb{R} \\ &\iff g(v) \approx g(v'). \end{aligned} \tag{S.1}$$

Now fix any proper u-upper set $\mathcal{U}$ in $\mathcal{V}$, and let $\mathcal{U}^* = g(\mathcal{U} \cap \mathcal{V}_s)$. To see that $\mathcal{U}^*$ is a u^*-upper set, fix any $v^* \in \mathcal{U}^*$ and $v^{*\prime} \in \mathcal{V}^*$ with $v^{*\prime} \gg_{u^*} v^*$. We need to show that $v^{*\prime} \in \mathcal{U}^*$. Let $v = g^{-1}(v^*) \in \mathcal{U} \cap \mathcal{V}_s$. Note that we cannot have $v^* \approx -u^*$, as this would imply by Equation (S.1) that $v \approx g^{-1}(-u^*) = -u$, which would in turn imply by the

[53] A more detailed argument is as follows: Define $h : \mathbb{R}^{n+2} \to \mathcal{V}_s$ by $h(\alpha) = \alpha_1 v_1 + \cdots + \alpha_n v_n + \alpha_{n+1} u + \alpha_{n+2}\mathbf{1}$ and define $h^* : \mathbb{R}^{n+2} \to \mathcal{V}_s^*$ by $h^*(\alpha) = \alpha_1 v_1^* + \cdots + \alpha_n v_n^* + \alpha_{n+1} u^* + \alpha_{n+2}\mathbf{1}^*$. By the linear independence of these sets of functions, both h and h^* are bijections. It is trivial that both functions are continuous, and by Aliprantis and Border (2006, Corollary 5.24) both h^{-1} and h^{*-1} are also continuous. Note that $g = h^* \circ h^{-1}$ and $g^{-1} = h \circ h^{*-1}$, and hence these functions are continuous.

definition of a u-upper set that $\mathcal{U} = \mathcal{V}$, contradicting our assumption that $\mathcal{U}$ is a proper subset of $\mathcal{V}$. Therefore, there exists some $\alpha \in [0, 1]$ such that

$$v^{*\prime} \approx \alpha u^* + (1 - \alpha)v^*.$$

Thus there exist $a > 0$ and $b \in \mathbb{R}$ such that

$$v^{*\prime} = a\alpha u^* + a(1 - \alpha)v^* + b\mathbf{1}^*.$$

Let

$$v' = a\alpha u + a(1 - \alpha)v + b\mathbf{1}.$$

Clearly $v' \in \mathcal{V}_s$. Moreover, since $v' \gg_u v$ we have $v' \in \mathcal{U}$. Thus $v' \in \mathcal{U} \cap \mathcal{V}_s$, which implies $v^{*\prime} = g(v') \in \mathcal{U}^*$.

(4): We can treat a lottery $p \in \Delta(C^*)$ as a measure defined only on the space C^*, or we treat this as a lottery in $\Delta(C)$ that assigns probability zero to the set $C \setminus C^*$. Thus we will abuse notation slightly and evaluate the lotteries $p \in \Delta(C^*)$ using both functions in $\mathcal{V}^*$ and functions in $\mathcal{V}$. Note that for any $v \in \mathcal{V}_s$, $v(p) = v^*(p)$ for $v^* = g(v) \in \mathcal{V}_s^*$. Therefore, for any $x \in \mathcal{K}(\Delta(C^*))$,

$$\begin{aligned}
U_i^*(x) &= \int_{\mathcal{V}^*} \max_{p \in B_{v^*}(x)} u^*(p)\, d\mu_i^*(v^*) \\
&= \int_{\mathcal{V}_s^*} \max_{p \in B_{v^*}(x)} u^*(p)\, d(\mu_i \circ g^{-1})(v^*) && \text{(definition of } \mu_i^*) \\
&= \int_{\mathcal{V}_s} \max_{p \in B_{g(v)}(x)} u^*(p)\, d\mu_i(v) && \text{(change of variables)} \\
&= \int_{\mathcal{V}_s} \max_{p \in B_v(x)} u(p)\, d\mu_i(v) \\
&= U_i(x).
\end{aligned}$$

Thus U_i^* is the restriction of U_i to $\mathcal{K}(\Delta(C^*))$. Also, note that μ_i^* is nontrivial (i.e., assigns probability zero to the set of constant functions) since

$$\mu_i^*(\{\alpha\mathbf{1}^* : \alpha \in \mathbb{R}\}) = \mu_i(g^{-1}(\{\alpha\mathbf{1}^* : \alpha \in \mathbb{R}\})) = \mu_i(\{\alpha\mathbf{1} : \alpha \in \mathbb{R}\}) = 0,$$

by the nontriviality of μ_i. Hence (u^*, μ_i^*) is a random Strotz representation of $\succsim_i^*$. ■

We now prove that $\mu_1 \gg_u \mu_2$. By assumption, $\succsim_2$ is more temptation averse than $\succsim_1$. Thus for any menu x and lottery p, $\{p\} \succ_1 x$ implies $\{p\} \succ_2 x$. This implies a fortiori that the same condition must hold for lotteries and menus of lotteries with support in C^*, and hence $\succsim_2^*$ is more temptation averse than $\succsim_1^*$, where $\succsim_i^*$ is defined as in part 4

of Lemma S.2. Since C^* is finite and (u^*, μ_i^*) represents $\succsim_i^*$ for $i = 1, 2$, Theorem S.3 implies that $\mu_1^* \gg_{u^*} \mu_2^*$.

Now fix any u-upper set $\mathcal{U}$ in $\mathcal{V}$. If $\mathcal{U} = \mathcal{V}$, then trivially $\mu_1(\mathcal{U}) = \mu_2(\mathcal{U}) = 1$. Otherwise, by part 3 of Lemma S.2, $g(\mathcal{U} \cap \mathcal{V}_s)$ is a u^*-upper set in $\mathcal{V}^*$ and therefore

$$\mu_1(\mathcal{U}) = \mu_1^*(g(\mathcal{U} \cap \mathcal{V}_s)) \geq \mu_2^*(g(\mathcal{U} \cap \mathcal{V}_s)) = \mu_2(\mathcal{U}),$$

where the equalities follow from part 2 of Lemma S.2 and the inequality follows from $\mu_1^* \gg_{u^*} \mu_2^*$. Since this is true for any u-upper set $\mathcal{U}$, conclude that $\mu_1 \gg_u \mu_2$.

S.2.2 Necessity: less u-aligned $\Longrightarrow$ more temptation averse

In this section we prove that the more temptation averse comparative is implied by $\mu_1 \gg_u \mu_2$. It is worth noting that the proof of this direction does not rely on the assumption that these measures have finite-dimensional support.

The following preliminary result will be useful.

Lemma S.3. *Let u, v, v' be expected-utility functions defined on $\Delta(C)$, and suppose $v \gg_u v'$. Then for any menu x,*

$$\max_{p \in B_v(x)} u(p) \geq \max_{q \in B_{v'}(x)} u(q).$$

Proof. If $v' \approx -u$, then for any menu x,

$$\max_{q \in B_{v'}(x)} u(q) = \min_{q \in x} u(q) \leq u(p), \quad \forall p \in x.$$

In particular,

$$\max_{q \in B_{v'}(x)} u(q) \leq \max_{p \in B_v(x)} u(p).$$

If we do not have $v' \approx -u$, then $v \gg_u v'$ implies $v \approx \alpha u + (1-\alpha)v'$ for some $\alpha \in [0, 1]$. First, consider $\alpha = 0$. In this case, $v \approx v'$. Therefore $B_v(x) = B_{v'}(x)$, which implies

$$\max_{p \in B_v(x)} u(p) = \max_{q \in B_{v'}(x)} u(q).$$

Finally, consider the case of $\alpha > 0$. Note that for any menu x and any $p \in B_v(x)$ and $q \in B_{v'}(x)$,

$$\alpha u(p) + (1-\alpha)v'(p) \geq \alpha u(q) + (1-\alpha)v'(q) \quad \text{and} \quad v'(q) \geq v'(p).$$

Since $\alpha > 0$, these inequalities imply $u(p) \geq u(q)$. Therefore,

$$\max_{p \in B_v(x)} u(p) \geq \max_{q \in B_{v'}(x)} u(q),$$

as claimed. ■

Suppose (u, μ_1) and (u, μ_2) are random Strotz representations of $\succsim_1$ and $\succsim_2$, and suppose $\mu_1 \gg_u \mu_2$. Fix any menu x, and let $[a, b] = u(x)$. Define $f_x : \mathcal{V} \to [a, b]$ by

$$f_x(v) = \max_{p \in B_v(x)} u(p).$$

By Lemma S.3, $v \gg_u v'$ implies $f_x(v) \geq f_x(v')$. Therefore, for any $\alpha \in [a, b]$ and $v \gg_u v'$,

$$v' \in f_x^{-1}([\alpha, b]) \iff f_x(v') \geq \alpha \implies f_x(v) \geq \alpha \iff v \in f_x^{-1}([\alpha, b]).$$

Thus $f_x^{-1}([\alpha, b])$ is a u-upper set. Therefore,

$$\mu_1(f_x^{-1}([\alpha, b])) \geq \mu_2(f_x^{-1}([\alpha, b])).$$

Define distributions $\eta_i^x \equiv \mu_i \circ f_x^{-1}$ on $[a, b]$ for $i = 1, 2$. By the preceding arguments, η_1^x first-order stochastically dominates η_2^x. Therefore, by the change of variables formula,

$$U_1(x) = \int_{\mathcal{V}} f_x(v)\, d\mu_1(v) = \int_a^b \alpha\, d\eta_1^x(\alpha) \geq \int_a^b \alpha\, d\eta_2^x(\alpha) = \int_{\mathcal{V}} f_x(v)\, d\mu_2(v) = U_2(x).$$

Since this is true for every x, and using the fact that $U_1(\{p\}) = U_2(\{p\})$ for any lottery p, it follows immediately that $\succsim_2$ is more temptation averse than $\succsim_1$.