Self-Control through Second-Order Preferences

Klaus Nehring
University of California, Davis

First Version:
September 11, 2006

Abstract

We propose to model the exercise of self-control as the second-order choice of one's own choice dispositions (first-order preferences over outcomes). This choice is governed by second-order preferences over first-order preferences and final outcomes. Specifically, the paper studies the revealed preference implications of the second-order preference model for ex-ante choices among opportunity sets in the abstract, non-probabilistic framework of Kreps (1979). While the implications of the general, unstructured SOP model turn out to be weak, additional restrictions on the relation between ex-post and ex-ante preferences over outcomes entail behavioral implications that distinguish the SOP model from other models of temptation and from other explanations of negative option value such as deliberation costs or unresolved value-conflicts/regret.

1. INTRODUCTION

In traditional conceptions of rational choice, additional options can never be harmful since the agent is always free not to choose them. But it has been recognized since ancient times that agents sometimes deliberately choose to reduce their options, for example by having themselves tied to the mast of a ship in order to prevent themselves from jumping the ship when exposed to the song of the sirens. Such precommitment is naturally (but not uncontroversially) viewed as the rational management of one's own perceived irrationality.

After being introduced into economics in the classical contribution by Strotz already in 1955, this theme has lead a surprisingly marginal existence for about 40 years, possibly because it had been perceived as empirically atypical while perhaps intriguing philosophically. Yet this perception has changed dramatically with the "decade of behavioral economics", as issues of self-management have become the topic of a rich, diverse and rapidly growing literature. Indeed, once agents are perceived to frequently not act in their own best interests, this forces the question of how they deal with their own irrationality squarely on the table.

There are two fundamental strategies of such self-management in the context of dynamic choices: the agent may change his future choice sets or his future choice dispositions. We will refer to these as the strategies of "precommitment" and "self-control", respectively. While the great majority of the literature following Strotz restricted attention to the strategy of precommitment, Thaler and Shefrin (1981) formulated the first economic model of self-control, and the seminal contribution by Gul and Pesendorfer (2001, henceforth: GP) provided the first rigorous decision-theoretic treatment. As noted in GP, the agent's exercise of self-control is revealed in his dynamic choice behavior, and more specifically in his choice among choice sets. To illustrate, think of Ulysses, approaching the sirens but not yet able to hear them, as choosing which choice set he will face when passing the them. In the original story involving pre-commitment, Ulysses prefers the choice set of being tied without choice (the choice set {Tied}) to the unconstrained choice set {Tied,Jump}; in view of his certain expectation that he would choose to jump off board in the latter case, he presumably would have ranked that set indifferent to leaving himself with no alternative but to jump (the set {Jump}), yielding the preference pattern

$$\{Tied\} \succ \{Tied, Jump\} \sim \{Jump\}.$$

By contrast, had Ulysses been more strong willed, or had the siren's song been less tempting, he would not have needed to take the drastic measure of asking to be tied to the mast; instead, he could

have roamed freely on board, relying on the force of his will-power to resist the sirens' temptation. Even so, the mere availability of the option to jump would have made him worse off, since it would have forced him to exercise his will-power.[1] The exercise of self-control is thus characterized by the distinct preference pattern

$$\{Roam\} \succ \{Roam, Jump\} \succ \{Jump\}.$$

While self-control shares with pure precommitment the potential undesirability of additional options, preferences based on self-control are harder to understand since the desirability of a choice set is no longer determined by the desirability of the ultimately chosen alternative only, but also by the utility costs arises from the non-choosing of the others.[2] To achieve a satisfactory decision-theoretic understanding of self-control, that is: of self-control preferences over menus, two issues need to be addressed. First, which patterns of menu choices can be explained in terms of self-control, and which cannot? What is the behavioral signature of self-control driven preferences? Second, is it possible to give a unified, workably general account of self-control preferences?

To answer these questions, we propose in this paper an account of "self-control as second-order preference maximization". Its starting point is the abstract, skeletal notion of self-control as an unobserved intra-psychic action that influences the agent's choice dispositions at the moment of final ("ex-post") choice. This action occurs at some time ("ex interim") between the ex-ante choice *of* the menu and the ex-post choice *from* the menu. What matters about the interim action are the induced choice dispositions; we take these to be representable in conventional terms by a preference ordering. The agent exercises self-control ex interim by choosing among "extended outcomes" that consist of a physical outcome together with the preference ordering that was chosen to achieve it. The interim choice maximizes a "second-order preference" (SOP) ranking over extended outcomes. Ex ante, the agent chooses the menu that offers the best extended outcome ex interim. The basic modeling idea is not really new; it is essentially a one shot version of the doer-planner model of Shefrin and Thaler (1981) and the dual self model of Fudenberg and Levine (2005).

While the conception of self-control as second-order preference maximization involves no logical paradox[3], it invites to be supplemented by some psychological or even neuro-scientific account that

[1] Or so the standard, somewhat narrow account goes. He might, however, have been vainglorious or reckless enough to prefer keeping the fatal option available in order to demonstrate his strength of character.

[2] As in GP, think of pure commitment preferences as a limiting case of infinitely costly self-control or "overwhelming temptation".

[3] Such as postulating two preference relations describing (different parts of) the agent at the same time (as suggested

explains the preconditions and mechanisms of such intra-psychic causation. Here, the recently influential distinctions between affective and deliberative and between automatic and controlled process are very promising. Indeed, these distinctions have already been explicitly invoked in support of an SOP-like modelling of self-control by Fudenberg and Levine (2005) as well as in more detail by Benhabib and Bisin (2004) and Loewenstein and Donoghue (2005).[4] These contributions have also demonstrated the economic relevance of SOP models of self-control by providing applications to a wide range of economic situations. However, none of these have provided decision-theoretic foundations, the goal of the present paper.

To do so, we study the implications of the second-order preference model for ex-ante choices among opportunity sets ("menus") in the abstract, non-probabilistic framework of Kreps (1979). The revealed preference implications of the general, unstructured SOP model turn out to be crisp, but weak: the characterizing condition called "Upper Boundedness" simply says that the union of two menus can never be superior to both of them. This condition corresponds, in fact, to exactly one half of GP's central Set Betweenness axiom.[5] Its bite is mainly to exclude menu preferences based on flexibility and temptation uncertainty.

The implications of the general SOP model are weak due to the absence of any restrictions on how the agent may override ex-ante his (endogenous) ex-post preferences. For example, the unstructured SOP model allows outcomes to be ranked ex-ante always in exactly contrary to how the agent would rank these outcomes ex-post, based on his endogenous ex-post preferences. Such second-order preferences may induce menu preferences in which additional options are always harmful – menu prefernences that hardly correspond to a sensible notion of self-control.

We argue that this apparent underdetermination of the behavioral content of the SOP model can be overcome by imposing further structure on second-order preferences that capture basic features of an intuitive, pre-formal notion of self-control. This leads to two refinements of the basic model, "second-order preferences with self-management" and "second-order preferences with self-command". The latter more restrictive class of *self-command* preferences assumes that ex-ante outcome valuations are independent of ex-post preferences; in this case, optimal self-control amounts to optimizing the trade-off between achieving desirable outcomes and minimizing expenditure of will-power. Self-

for example by Shefrin and Thaler (1981))

[4] While Bernheim and Rangel's (2004) distinction between cold and hot modes also appeals to such a distinction, they do not model self control in the sense of the present paper since in their model the ex-post choice disposition is determined by external, random cues rather than the agent himself.

[5] Downward Monotonicity was first isolated by Dekel et al. (2005) under the name of "Positive Betweenness".

command preferences are characterized by Upper Boundedness plus a property called "Singleton Monotonicity" which requires that an agent is never made worse off by the addition of an alternative that is superior (as a singleton choice set) to all alternatives in the given menu.

Self-command preferences are restrictive, however, for frequently the agent will take into account ex-ante how the outcomes are ranked ex-post. For example, in a story of optimal wishful thinking based on endogenously chosen beliefs broadly along the lines of Brunnermeier and Parker (2005), the agent's ex-ante well-being would depend on the anticipatory utility derived from the endogeneously chosen wishful beliefs. Even though the agent may know at some level that the chosen ex-post beliefs are not rationally justifiable, he may sensibly take the "felicity" derived from these beliefs to be real and thus matter ex-ante. To accommodate such situations, the broader class of *self-management preferences* allows ex-ante preferences to positively reflect ex-post preferences; ex-post preferences can be overruled ex-ante, but only in a consistent way. The main result of the paper (Theorem 12) characterizes self-control preferences in terms of Upper Boundedness plus a property called Limited Temptation. Limited Temptation is intermediate in strength between Singleton Monotonicity on the one hand and, on the other, the requirement that every menu be at least as desirably as the worst alternative that it contains.

Comparison to the Literature

The seminal contribution to the decision theoretic literature on self-control is the already mentioned paper by Gul and Pesendorfer (2001, "GP"). GP provides an elegant and highly parsimonious axiomatic model that has stimulated a sizeable and growing follow-up literature, both axiomatic and applied. In order to achieve this parsimony and simplicity of functional form, GP deliberately sacrificed generality. It is thus of obvious interest to investigate what forms self-control driven choice behavior can take more generally. Other papers addressing the question of generality– all couched as generalizations of the GP model– are Dekel, Lipman and Rustichini (2005), Noor (2006) and Chatterjee and Krishna (2005). The first two will be discussed below, the last is less germane here as it focus on temptation uncertainty.

The present paper departs from GP – and indeed from the entire axiomatic literature to date – in its emphasis on self-control rather than temptation. In the GP model, self-control enters as an interpretation of the obtained functional form. In point of fact, on a more straightforward interpretation, GP menu utilities are simply the sum of a positive (ordinary) indirect utility and

a negative (temptation) indirect utility; in other words, GP preferences *could* be interpreted as if the fact of being tempted was a bad in itself, without any role for self-control. This does not mean that a fleshed-out self-control interpretation of the GP model is not possible or appropriate; indeed, we point out below in section 3.2 that GP menu preferences can be derived from a second-order preferences with self-command with a rather special but fairly natural structure.

GP's emphasis on temptation rather than self-control limits the ability of the GP model to accommodate important aspects of the exercise of self-control. First of all, one very robust feature of optimal self-control will be its responsiveness to the incentives for self-control, i.e. to the gains in outcome utility relative to the effort of will-power. As we show in an adaptation of the model by Benhabib and Bisin (2004), this will typically lead to ex-post choice behavior that violates standard context-independence (or choice "consistency") conditions. By contrast, an important part of the simplicity of the GP model is the implicit assumption of context-independent ex-post choice[6]. In a related vein, Fudenberg and Levine (2005) have observed that cognitive load effects suggested in the psychological and experimental literature such as Shiv and Fedorikhin (1999) and Muraven and Baumeister (2000) lead to failures of context-independence.

Second, if one embeds the GP model within the SOP model, it turns out that the ex-post preferences implicit in GP menu preferences *must* violate standard well-behavedness assumptions such as preference convexity or the expected-utility hypothesis; see section 3.5. We interpret this finding as suggesting that the GP model implies a particular kind of self-control through *desire repression* as opposed to self-control through *desire modification.*

Third, being a model of self-command, the GP does not allow ex-ante outcomes to depend on ex-post preferences, as discussed above.

Both Dekel et al. (2005) and Noor (2006) generalize the GP model, the first by introducing multiple temptations, and the second by making the strength of temptations menu-dependent. Both contributions retain the emphasis on temptation rather than self-control, and end up characterizing classes of menu preferences that are very different from the class of self-control and self-command preferences at the center of this paper. Along with GP, they also assume that the agent has well-defined preferences over menus of lotteries rather than menus of abstract alternatives.

Dekel et al. (2005) retain GP's Independence axiom but weaken their Set Betweenness axiom;

[6] More rigorously, of context-independence of the ex-post choice behavior suggested by the functional form and appealed to in its interpretation. This implied context-independence comes to the fore axiomatically in their recent 2006 paper which presents a lottery-free counterpart to the original GP paper.

indeed, in the case of deterministic temptation on which we shall focus as it overlaps with the goals of the present paper,[7] they weaken Set Betweenness to the Upper Boundedness axiom described above. The Independence axiom applied to menus of lotteries is very strong; among other things, it implies context-independence of ex-post choices, which, as mentioned above, is severely restrictive from a self-control perspective. Indeed, as pointed out by Fudenberg and Levine (2005), if self-control is understood as a "missing action" that co-determines the realized value of a menu, the Independence axiom is problematic for essentially the same reason for which it is inapplicable in the case of ordinary lottery preferences with a missing action; see Machina (1984) and Mas-Collel, Whinston and Green (1995). As observed in section 6, multiple temptations preferences typically violate Singleton Monotonicity and Limited Temptation and, by consequence, do not admit a well-behaved self-control interpretation in the sense of the present paper.

Noor (2006) presents various examples to show that context-independence of implied choices is seriously restrictive. He thus drops Independence, and weakens Set Betweenness quite drastically. Noor's representation allows temptations to be context-dependent; while the resulting model is very flexible, it does not yield or suggest much structure since the context-dependence of temptations is left unexplained. By contrast, in the present paper, the context-dependence of choices is derived from the maximization of a context-independent second-order preference ordering. Menu preferences in his model may easily violate both Upper Boundedness and Limited Temptation.

At the methodological level, all of the above contributions characterize preferences over menus of lotteries, rather than generic, abstract "alternatives" as done here. Reference to lotteries is avoided in the present paper because lotteries are extraneous to the notion of self-control, and because the purpose of the present paper is to characterize the implications of conceptualizing self-control as second-order preference maximization in general. At a more pragmatic level, avoidance of any a priori structure on alternatives allows one to obtain results for (small) finite domains of menus. This enhances the testability of the model in practice, and also makes the assumption of a well-defined complete preference ordering over menus more plausible.[8]

The perhaps most troublesome feature of the SOP model (at the level of generality considered here) is the absence of interesting uniqueness properties; the interest in such properties presumably was a main driver behind the use of the lottery framework in the above contributions and the willingness

[7] Their paper is substantially more general by also allowing for uncertainty of the temptation.

[8] Gul-Pesendorfer (2006) axiomatize a finite, lottery-free counterpart to their earlier representation.

to impose sometimes very strong assumptions in that framework. We view the non-uniqueness of the representation not as fatal deficiency of the model but as an inescapble "fact of life", a (not inconsiderable) technical inconvience and an interesting research question. First and foremost, the appeal of the model does not come exclusively, or even primarily, from a representation theorem, but from the conceptually compelling view of self-control as an unobserved psychic action. It is thus no accident that versions of the SOP model have been employed before in mentioned contributions of Shefrin and Thaler (1981), Benhabib and Bisin (2004) and Fudenberg and Levine (2005) prior to any axiomatization.

The absence of useful uniqueness properties is due partly to the lack of structure on the alternatives and partly to the lack of structure imposed on second-order preferences. It will be an important topic for future research to explore to what extent interesting uniqueness results can be obtained in special cases. One interesting such special case would be based on assumption that all feasible preferences over lotteries have the expected-utility form. Preliminary investigations suggest that, at least in particular subcases, uniqueness is obtainable.

2. SECOND-ORDER PREFERENCES

The behavioral primitive is a ranking (weak order) $\succsim$ over a domain $\mathcal{M}$ of opportunity sets or "menus". Menus $A, B, ...$ are simply sets of alternatives taken from some finite ground set X. The domain $\mathcal{M}$ will generally be assumed to be *comprehensive,* that is: to contain all singletons and to contain any subset of any set it contains ($A \in \mathcal{M}$ implies $B \in \mathcal{M}$ for any $B \subseteq A$). Thus $\mathcal{M}$ might consist of all non-empty subsets of X, or of possibly much smaller subfamilies such as the set $\mathcal{M}_m$ of all menus of cardinality not exceeding m, with $m = 2$ or 3. This generality is helpful since much of the intuition and presumably practical testability will come from choices among menus of small cardinality.

The second-order preference (SOP) model explains menu preferences as follows. At date 2 ("ex post"), the agent chooses an alternative from the menu A based on the first-order preference (choice disposition) P over X; to ensure single-valued ex-post choices, P is assumed to be a linear order.[9]

The first-order preference ordering P, is itself chosen "ex interim", after the menu A had been chosen at date 1 ("ex ante"). The interim choice of P in turn is based on the agent's second-order

[9] A linear order is a transitive, complete and anti-symmetric relation, that is: a weak order with only trivial indifferences of the form xPx.

preference relation $\trianglerighteq$ over "extended outcomes" $(x, P) \in X \times \mathcal{L}(X)$, where $\mathcal{L}(X)$ is the set of linear orders on X. Thus, the agent cares about not only the ultimate physical outcome x, but also about the choice-disposition P he "chose" or "formed" in order to obtain that outcome. The relation $\trianglerighteq$ is assumed to be a weak order on $X \times \mathcal{L}(X)$ that is numerically represented by a "second-order utility function" V. Second-order strict preference and indifference will be denoted by $\triangleright$ and $\equiv$, respectively.

Ex-interim, when forming P, the agent anticipates that he will end up choosing the P-maximal alternative in A $P(A)$; as usual, $x = P(A)$ iff xPy for all $y \in A$. He thus forms P to obtain the best extended outcome $(P(A), P)$. Likewise, when choosing among menus ex-ante, the agent evaluates menus according to the best extended outcome $(P(A), P)$ they enable.[10] That is, the agent ranks menus according to

$$A \succsim B \text{ iff } \arg\max_{\trianglerighteq}\{(P(A), P) : P \in \mathcal{L}(X)\} \trianglerighteq \arg\max_{\trianglerighteq}\{(P(B), P) : P \in \mathcal{L}(X)\}. \tag{1}$$

Equivalently, there exists a menu-utility function $U : \mathcal{M} \to \mathbf{R}$ representing menu preferences $\succsim$ such that

$$U(A) = \max_{P \in \mathcal{L}(X)} V(P(A), P); \tag{2}$$

it is frequently instructive to break down (2) further and write

$$U(A) = \max_{x \in A} u(x, A), \tag{3}$$

with

$$u(x, A) = \max_{P : P(A) = x} V(x, P). \tag{4}$$

The expression (3) represents the menu ranking as an indirect utility based on the context-dependent $u(x, A)$, and (4) explains this context-dependence in terms of the "incentive compatibility constraint" $\{P : P(A) = x\}$. The difference $u(x) - u(x, A)$ can be thought of as the "cost of self-control" associated with implementing x; as expanding the set A will tighten incentive compatibility the constraint, this cost will weakly increase in A for fixed x.[11]

[10] The assumption that ex-ante and ex-interim choices are based on the same second-order preference ordering is restrictive; one can imagine situations (especially when there is a large temporal gap between the menu and the outcome choices) in which ex-ante the agent anticipates excessive or misguided self-control efforts ex-interim (e.g. out of compulsiveness). To incorporate this, one could allow for third- and higher-order preference orderings, at the price of significant losses in the tractability and explanatory content of the model.

[11] The cost of self-control $K(x, A) = u(x) - u(x, A)$ plays an important role in the exposition of Fudenberg-Levine (2005). However, in contrast to the present paper, the general model of FL (adapted to the present framework) allows

The scenario underlying self-control through second-order preferences is summarized by the following time line.

Time Line

Date	Choice of	Choice based on
Ex ante	A	$\trianglerighteq$
Ex interim	P	$\trianglerighteq$
Ex post	x	P

A few related pieces of notation will be helpful. For any $A \in \mathcal{M}$,

- the set of (ex-interim) *feasible extended outcomes* is denoted by

  $Y_A := \{(P(A), P) :, P \in \mathcal{L}(X)\}$;

- the set of (possibly) *chosen extended outcomes* is

  $H(A) := \arg\max_{\trianglerighteq} Y_A = \{(x, P) \in Y_A : (x, P) \trianglerighteq (x', P') \text{ for all } (x', P') \in Y_A\}$, and

- the set of (possibly) *ex-post chosen alternatives* is

  $C(A) := \{x : (x, P) \in H(A) \text{ for some } P \in \mathcal{L}(X)\}$.

Also, we will denote the restriction of $\succsim$ to singletons by $\succsim_1$; $\succsim_1$ describes the agent's *commitment preferences* which reflect his ex-ante valuation of outcomes in the absence of self-control issues.

To illustrate how second-order preferences rationalize menu preferences, consider the simplest instance of self-control, namely the menu preference

$$\{a\} \succ \{a, b\} \succ \{b\},$$

with $X = \{a, b\}$.

These preferences are rationalized by the following SOP,

$$(a, ba) \rhd (a, ab) \rhd (b, ba) \rhd (b, ab),$$

a particular first-order preference P being denoted by listing the alternatives in decreasing order (here and throughout the rest of the paper). In the above SOP, it takes effort reflected in reduced second-order utility to induce a first-order preference for a over b : $(a, ba) \rhd (a, ab)$; moreover, this effort

V also to depend directly on the choice set A. As a result, the self-control cost function K in FL does not need to satisfy this monotonicity property, nor any further restrictions entailed by the functional form (4).

is worthwhile since $(a, ab) \rhd (b, ba)$. Thus, the extended outcome choices are $H(\{a\}) = \{(a, ba)\}$, $H(\{b\}) = \{(b, ba)\}$, and $H(\{a, b\}) = \{(a, ab)\}$. Since $H(\{a\}) \rhd H(\{a, b\}) \rhd H(\{b\})$, these lead to the menu preferences $\{a\} \succ \{a, b\} \succ \{b\}$.

The ex-post choice function C induced by a given second-order preference ordering $\trianglerighteq$ can be interpreted in two ways. First, and most straightforwardly, C can stand for the agent's actual ex-post choices. As such, C is behaviorally observable, and could thus legitimately have been included among the primitives. We have abstained from doing so to make the model directly comparable to the existing decision-theoretic literature on self-control, but also because under natural assumptions the implied choices are uniquely determined by menu preferences, as we will show in a companion paper.[12]

Alternatively, C (and H) could be interpreted as the agent's point expectations governing his interim preference formation. These may differ from his actual choices if the agent is mistaken about the ex-post choice dispositions that in fact result from a particular internal self-control action. For example, the agent may naively believe that he would act in line with his ex-ante preferences over outcomes (and thus fail to exhibit any desire to precommit), only to succumb to temptation once faced with the actual choice. While the first interpretation assumes that the agent is "sophisticated" in correctly forecasting the effect of his self-control actions on the resulting choice dispositions, the second interpretation is neutral on how sophisticated or naive the agent in fact is; for models of dynamic choice that allow for partially sophisticated agents, see O'Donoghue and Rabin (1999, 2001).

Note that, given a context-dependent indirect utility representation (3) and (4), C is given as

$$C(A) = \arg\max_{x \in A} u(x, A); \tag{5}$$

the context-dependence of u suggests that one should not expect C itself to satisfy standard choice-consistency/context-independence conditions such as IIA[13]. And, indeed, we will encounter many natural violations of IIA in the sequel. We note that the implied choices in the GP model, as well as in Dekel et al. (2005) and Noor (2006), also satisfy (5).

[12] In particular, it can be shown that if the menu preference satisfyies Limited Temptation and is derived from a linear second-order preference ordering, the resulting choice-function is uniquely determined. As the role of linearity throws up additional conceptual and technical issues, we do not include this material here.

[13] IIA says that if an alternative is chosen in a larger set, then it must also be chosen in any smaller set in which it is feasible; formally, for any x, A, B such that $B \subseteq A : x \in C(A) \cap B$ implies $x \in C(B)$. We will identify context-dependence with IIA throughout.

As a matter of psychological technology, it is clear that an agent will typically not be able to get himself to form arbitrary ex-post choice-dispositions, no matter which self-control tactics he engages in. Infeasibility of a choice disposition is behaviorally equivalent to assuming it to be prohibitively costly. In terms of the SOP ordering $\trianglerighteq$, this is captured by deeming the first-order preference P *infeasible* if there exists $Q \in \mathcal{L}(X)$ such that, for all $x, y \in X$, $(x, Q) \triangleright (y, P)$; we will denote the complementary set of feasible preferences by $\mathcal{P}$.

A limiting case obtains when the agent cannot influence his future choice dispostions at all which is captured by the existence of a unique feasible preference ($\mathcal{P} = \{P\}$). In this case, menus are ranked according to

$$A \succsim B \text{ whenever } \{P(A)\} \succsim \{P(B)\}.$$

We will call such menu preferences "menu preferences *without self-control* ".[14] In the special case in which ex-ante and ex-post outcome preferences agree ($P = \succsim_1$), menus are ranked according to their indirect utility.

3. EXAMPLES

To illustrate the explanatory power of the SOP model, we will now present a number of examples some of which are closely related to models in the literature. To focus on the central trade-off between achieving ex-ante optimal outcomes and economizing on self-control costs, all of them will adopt the following additively separable functional form:

$$V(x, P) = u(x) - k(P), \tag{6}$$

for appropriate functions $u(.)$ and $k(.)$ such that $\min_{P \in \mathcal{L}(X)} k(P) = 0$. The disutility $k(P)$ can be interpreted as the "cost of will-power" of adopting P which is traded off against the direct utility of the physical outcome $u(x)$; infeasible preferences are those with $k(P) = \infty$. If there is a unique cost-minimal preference P, this preference can be interpreted as the choice disposition resulting from the agent's automatic decision-processes in the absence of any self-control efforts, and will be referred to as the agent's *default* preference ordering D. We will show that even in very simple cases, (6) leads to interesting phenomena that violate basic assumptions of the GP model.

[14] GP refer to this as "temptation without self-control" or "overwhelming temptation".

3.1. Fixed costs of cognitive control

The simplest specialization of the additively separable SOP model (6) results from the existence of only two feasible preferences, $\mathcal{P} = \{Q, D\}$, with Q ranking alternatives according to their ex-ante utility u, $k(D) = 0$, and $\gamma = k(Q) > 0$ denoting the fixed cost of acting according to Q; this is essentially a static version of the model of Benhabib-Bisin (2004) who applied it a dynamic consumption-savings problem. Benhabib-Bisin provide a detailed motivation inspired by neuroscience, interpreting γ as the fixed cost of switching from automatic to controlled cognitive processes; with a somewhat different spin, one can interpret γ as the fixed cost of breaking the "hot" affect provoked by the situation and acting with a cold head.

To illustrate the behavioral implicatins of the model, suppose that the agent has a mild problem with excessive alcohol consumption, without being genuinely addicted. Let $X = \{0, ..., L\}$, with x denoting the amount of alcohol (in milliliter) consumed over a specified time interval (e.g. an evening). D ranks the elements of X in increasing order: that is, the agent's default tendency is to consume as much alcohol as possible. On the other hand, Q and $u(.)$ rank X in decreasing order (with $u(x) = -x$): ex ante (before the party has begun), the agent views alcohol consumption as a bad. It takes him $\gamma > 1$ utiles of effort to act on the ex-ante preferences. This results in the following choice behavior: if faced with a choice among similar amounts of alcohol (if $\max A - \min A < \gamma$), the agent lets himself go and follows his default penchant for more alcohol; on the other hand, he will never end up drinking more than γ ml more than unavoidable.

Formally, ex post choices are given by the choice function

$$C(A) = \begin{cases} \max A & \text{if } \max A - \min A < \gamma \\ \min A & \text{if } \max A - \min A \geq \gamma \end{cases},$$

and menus are ranked according to

$$U(A) = \max(-\max A, -\min A - \gamma).$$

Note in particular that ex-post choices are context-dependent: observed choices from menus with small stakes ($\max A - \min A < \gamma$) appear to reveal a preference for ever larger amounts of alcohol; but this preference is contradicted by his self-restraint when the stakes are sufficiently large ($\max A - \min A \geq \gamma$).[15]

[15] In the GP model which builds in context-independent ex-post choice, this is replicated in mistaken inferences on menu-preferences: observing the menu preferences $\{x\} \succ \{x, x+1\} \sim \{x+1\} \succ \{x+1, x+2\} \sim \{x+2\} \succ ...$, , these

The message of this example is robust. It is a fundamental intuition about optimally economizing on self-control that the amount of self-control exerted will increase with the stakes (in terms of outcome utility gained). With the exertion of more self-control, the ex-post choice dispositions will be increasingly brought into line with ex-ante preferences over outcomes. Due to this systematic shift of the adopted ex-post preference with the choice set, one would *expect* the induced choice-function to exhibit context-dependence.

3.2. A non-linear generalization of the GP model

Heuristically, it can be instructive to think of the cost of will-power in forming P as being determined by the "distance" of P from the default preference D, with the distance measuring the extent to which D must be distorted to obtain P. Formally, this involves writing

$$k\left(P\right) = \phi\left(d(P, D)\right). \tag{7}$$

A natural specification of such a metric d yields a non-linear generalization of the GP model. The metric relies on a fixed cardinal utility function u_D representing the default preference D. This allows one to measure the distance of P from D as the minimal amount by how much outcome utilities must change in order to transform D into P. Formally, let

$$d(P, D) := \inf_{\widetilde{u}:\widetilde{u} \text{ represents } P} \sup_{x \in X} |\widetilde{u}(x) - u_D(x)|. \tag{8}$$

It is easily checked that (6), (7), and (8) lead to a context-dependent utility

$$u(x, A) = u(x) - \phi\left(\max_{y \in A} u_D(A) - u_D(x)\right). \tag{9}$$

Equation (9) obtains since, in order to make x top-ranked in A, under (8) the perceived utility of x needs to be lifted vis-a-vis the best default choice $\arg\max_{y \in A} u_D(A)$ by at least $u_D(x) - \max_{y \in A} u_D(A)$ (+ arbitrarily small ϵ).

If ϕ is the identity, (9) yields the GP representation, with ex-post choices maximizing

$$u(x) + u_D(x),$$

satisfying context-independence. As pointed out by Fudenberg-Levine (2005), with non-linear ϕ, this context-independence is lost. They argue in particular that the notion of will-power as using

rankings would lead one to infer that menus are ranked in inverse order of their maximal element, i.e. according to $U(A) = -\max A$, which is wide off the mark.

scarce cognitive resources suggests that ϕ should be convex rather than linear. Note also that the fixed-costs model can be viewed as a special, limiting case of (9) by taking

$$\phi(v) = \begin{cases} \gamma & \text{if } v > 0, \\ 0 & \text{if } v = 0. \end{cases}$$

We view the SOP model's ability to embed the GP in a fairly simple and transparent way as an indication of its unifying potential.[16]

While the fixed-costs model and the non-linear extension of the GP model entail interesting departures from the original GP model, these departures remain fairly modest, since they retain the key simplifying feature that the context-dependent utility of an alternative depends on the menu only via the "most tempting" alternative(s) $\arg\max_{y \in A} u_D(A)$.[17] By consequence, such menu-preferences continue to satisfy GP's central Set Betweenness axiom.[18] This axiom has two parts which we shall refer to as Upper and Lower Boundedness.[19]

Axiom 1 ***(Upper Boundedness)*** *For all $A, B \in \mathcal{M}$: $A \cup B \precsim A$ or $A \cup B \precsim B$.*

Axiom 2 ***(Lower Boundedness)*** *For all $A, B \in \mathcal{M}$: $A \cup B \succsim A$ or $A \cup B \succsim B$.*

The first condition is fundamental; it reflects the negative context-dependence that is built into the SOP model and will turn out to characterize it; cf. Theorem 3 below. By contrast, the second condition is quite special and may easily be violated, as the following two examples will show.

3.3. Multiple temptations

Say that an alternative x *tempts in the menu* A if $A \succ A \backslash \{x\}$, that is: if the agent would want to commit not to choose x from A. Lower Boundedness is easily seen to imply that every menu contains at most one tempting alternative. This is a key to the simplification of the GP model, but it is clearly restrictive as illustrated by the following example.

[16] Admittedly, this embedding is at this point heuristic and lacks a rigorous articulation, for example in terms of axiomatic conditions on second-order preferences. Such an articulation would presumably have to be formulated in a lottery framework as in the original GP paper.

[17] This is Assumption 5 ("Opportunity Based Cost of Self-Control") in Fudenberg-Levine (2005).

[18] See Fudenberg-Levine (Theorem 5).

[19] DLR, which were the first to formally separate these two conditions, refer to them as Positive and Negative Betweenness, respectively. We chose to depart from their nomenclature, since one half of a "betweenness" condition has no betweenness left in it at all, and since the positive vs. negative distinction is specific to their model.

Consider the following pair of preferences over dessert menus based on the alternatives no dessert, light dessert, heavy dessert;[20] here and elsewhere, starred alternatives indicate the choices implied by the subsequent explanation of the preferences:

$$\begin{aligned} \{\text{none}^*, \text{heavy}\} &\succ \{\text{none}, \text{light}^*\} \succ \{\text{none}, \text{light}^*, \text{heavy}\} \text{ , and} \\ \{\text{none}, \text{light}^*\} &\succ \{\text{none}, \text{light}^*, \text{heavy}\}. \end{aligned}$$

Lower Boundedness is violated since the two alternatives "light" and "heavy" are both tempting in the menu {none, light, heavy}. A natural story is the following. Our agent, a weight-watching dessert lover, is tempted by a heavy desert, but can resist this temptation easily as he is just too aware of the conflict with his long-term interest in weight-control. Thus the menu {none, heavy} is just slightly worse than the ability to commit to no dessert at all (the menu {none}). On the other hand, the temptation by a light dessert is a strong one: the 'voice of reason' speaks only mutedly since a light dessert is not so bad, and our agent is a dessert lover after all. Thus, if the light dessert is available, he will take it, leaving him worse off than had he been tempted by a heavy dessert only; in partiuclar, {none, heavy}$\succ${none, light}. However, if both desserts are available, he will now need to exercise some will-power not to fall for the heavy dessert, leading to the menu-preference {none, light}$\succ${none, light, heavy}.

This story can be captured in the additively separable SOP model as follows. There are two feasible preferences Q and D over the alternatives $\{n, \ell, h\}$, with $\ell Q n Q h$ and $h D \ell D n$. The outcome utilities are $u(n) = 4$, $u(\ell) = 2$, $u(h) = 0$, and preference disutilities are $k(Q) = 1$ and $k(D) = 0$. One the easily computes

$$U(\{n, h\}) = V(n, Q) = 3 > U(\{n, \ell\}) = V(\ell, D) = 2 > U(\{n, \ell, h\}) = V(\ell, Q) = 1,$$

rationalizing the above preferences.

3.4. Multiple tasks

Consider an agent faced with a number of prima-facie unrelated choices each of which requires the exercise of self-control. To keep things simple, each of these choices will be binary. For example,

[20] The same preference pattern with a somewhat different story has been discussed before by Dekel et al. (2005). Indeed, GP already envisioned this and related violations of Set Betweenness quite clearly (GP, p. 1408-1409); they state that "we rule out these more elaborate formulations of temptation, as well as other deviations from the standard model, to stay close enough to the standard model so that the difference in behavior can be attributed solely to the presence of temptation."

the agent may be faced with the choice whether or not to keep his diet, whether or not to exercise, whether or not to keep moderation in drinking, etc. The choices are related indirectly through the use of "will-power" as general-purpose resource that is in limited supply and used up by the exercise of self-control. This notion of will-power as a limited resource has been proposed by Muraven, Tice and Baumeister (1998) and modeled in economics by Ozdenoren, Salant and Silverman (2006); see also Loewenstein and Donoghue (2005). It leads to the prediction that demands on self-control in certain tasks reduce the extent of self-control in others. Ozdenoren et al. argue that this prediction corresponds to stylized facts about people's observed life-style choices, in particular to the fact that much of the individual variation in health-related self-control behaviors concerns less the overall number of areas in which an individual exhibits self-control problems than the particular areas in which these show up. In contrast to the formalism but not the spirit of their model, the following model explicitly captures the notion of will-power being used up by *choices* involving self-control (e.g. of the choice to exercise rather than to enjoy leisure) rather than particular *activities* (e.g. the activity of exercising itself).

In this toy model, there is a set J of activities the agent can choose to engage in or not; this choice is simultaneous. An alternative is thus a vector of zeros and ones, $x \in X = \{0,1\}^J$. Commitment preferences are strictly monotone; in the absence of self-control problems, an agent is thus always better off doing the activity than not doing it. On the other hand, choosing to do this activity requires the expenditure of one unit of will-power. This can be modeled in the SOP framework as follows.

The intrinsic unrelatedness of the tasks is captured by assuming that all feasible preferences $P \in \mathcal{P}$ are weakly separable. That is, for all $P \in \mathcal{P}$, the following condition holds:

$$(1, x_{-j})P(0, x_{-j}) \text{ iff } (1, y_{-j})P(0, y_{-j}), \text{ for all } j \in J, x, y \in \{0,1\}^J.$$

This condition states that, for each feasible preference $P \in \mathcal{P}$, one can say whether or not the agent prefers to do a particular activity j irrespective of what else he does. The default preference is to be disposed against any particular activity. Acquiring a disposition in favor of an activity requires a unit of will-power; this leads to a cost of will-power function on $\mathcal{P}$ of the following form:

$$k(P) = \widetilde{k}\left(\#\{j : (1, y_{-j})P(0, y_{-j})\}\right),$$

where $y \in \{0,1\}^J$ is an arbitrary vector of activities.

The notion of will-power as a limited resource in fixed supply is captured by assuming in addition

that

$$\begin{aligned} \widetilde{k}(e) &= \begin{cases} 0 & \text{if } e \leq \overline{e}, \\ \infty & \text{if } e > \overline{e}, \end{cases} \\ \text{with } 1 &< \overline{e} < \#J. \end{aligned}$$

This case is of particular interest since it implies that all costs of self-control are opportunity costs. Note in particular that it implies that, for all pairs x, y, $\{x, y\} \sim \{x\}$ or $\{x, y\} \sim \{y\}$, a property which characterizes within the GP model the class of menu preferences without self-control.

The specified second-order preferences lead to menu preferences that violate Lower Boundedness and thereby Set Betweenness. In particular, since $1 < \overline{e}$, the agent can costlessly exercise self-control if faced with only a single activity choice at a time; with $\mathbf{1}$ denoting the vector $(1, .., 1)$, one thus has

$$\{\mathbf{1}\} \sim \{\mathbf{1}, (0, \mathbf{1}_{-j})\} \text{ for all } j. \tag{10}$$

On the other hand, if one enables the agent to opt out of any particular activity at the same time, the agent would lack the will-power to resist each time, thereby ending up worse off ex-ante:

$$\{\mathbf{1}\} \succ \{\mathbf{1}\} \cup \{(0, \mathbf{1}_{-j})\}_{j \in J}; \tag{11}$$

Clearly, (10) and (11) are inconsistent with a repeated application of Lower Boundedness. If $\overline{e} = \#J - \frac{1}{2}$, for example, each alternative in $A = \{\mathbf{1}\} \cup \{(0, \mathbf{1}_{-j})\}_{j \in J}$ is tempting in A except the alternative $\mathbf{1}$.

3.5. Self-Control by Modification of Desire vs. Self-Control by Repression of Temptation

The specialness of GP's approach emerges especially clearly in economic settings in which the space of alternatives is endowed with additional structure, especially linear structure. Thus, let X denote a finite subset of $\mathbf{R}^n$, and $\mathcal{M}$ a comprehensive set of menus in X.

For concreteness, consider an agent's choice of two-period consumption streams $x = (x_2, x_3)$; at date 1, the agent chooses a menu of consumption streams, and at date 2, he chooses a consumption stream over this and the following period. In this set-up, we can model the standard issue of present-bias: in the absence of self-control efforts, the agent's ex-post choice is characterized by a high degree of impatience, resulting in small savings and small future (date 3) consumption. Ex-ante, the agent prefers more future consumption and would thus like to commit to more saving. Alternatively, he might exercise self-control by forming more patient ex-post preferences.

This can be naturally modeled in the SOP approach as follows. Feasible preferences $P_\delta \in \mathcal{P}$ are parametrized by a discount factor $\delta \in [\delta_*, \delta^*] \subseteq [0,1]$, where δ_* denotes the "default" and δ^* the (ex-ante) "ideal" discount factor; the fixed temporal utility from consumption is given by a strictly increasing and strictly concave utility function $h : \mathbf{R}_+ \to \mathbf{R}$. So P_δ is given via its utility representation u_δ by

$$u_\delta(x) = h(x_1) + \delta h(x_2).$$

Second-order preferences are given by the second-order utility function

$$V(x, P_\delta) = u_{\delta^*}(x) - k(\delta),$$

where $k : [\delta_*, \delta^*] \to \mathbf{R}_+$ is non-decreasing in δ, with $k(\delta_*) = 0$ and $k(\delta) > 0$ if $\delta > 0$. For expositional specificity, we will assume that ex post in the absence of self-control the agent cares only about present consumption ($\delta_* = 0$), but is perfectly patient ex ante ($\delta^* = 1$). Consider the agent's preferences over menus composed of the consumption streams $a = (100, 100)$, $b = (150, 50)$, and $c = (200, 0)$. Clearly, $u_{\delta^*}(a) > u_{\delta^*}(b) > u_{\delta^*}(c)$. Thus, assuming the costs of will-power $k(.)$ to be sufficiently low, consumption streams with greater present consumption tempt those with less present consumption, but this temptation is resisted, i.e.

$$\{a\} \succ \{a, b\} \succ \{b\} \succ \{b, c\} \succ \{c\}. \tag{12}$$

Now compare the ranking of the missing sets $\{a, c\}$ and $\{a, b, c\}$ implied by this preference pattern within the SOP model to the ranking implied by the GP model. In the SOP model, it follows from the convexity of feasible preferences alone that

$$\{a, b\} \precsim \{a, b, c\} \precsim \{a, c\}, \tag{13}$$

hence that neither adding c to b nor substituting c for b can do any harm.

To see this, note that the assumption that the temptation of higher present consumption is resisted at $\{a, b\}$ (i.e. $\{a, b\} \succ \{b\}$) implies that the agent will chose a in $\{a, b\}$, together with an appropriate preference P implementing a, i.e. $(a, P) \in H(\{a, b\})$ with aPb. From the convexity of P, it follows that aPc, hence that the extended outcome (a, P) is feasible in $\{a, b, c\}$ and $\{a, c\}$ as well. By implication, in both $\{a, b, c\}$ and $\{a, c\}$ the agent is able to do at least as well as in $\{a, b\}$. Finally, Upper Boundedness allows one to infer that $\{a, b, c\} \precsim \{a, c\}$, completing the verification of (13).

Furthermore, in the SOP model, it will typically be the case that

$$\{a, b\} \sim \{a, b, c\} \prec \{a, c\};$$

indeed, in the present example, this will happen whenever will-power costs $k(\delta)$ are strictly increasing. For in this case the agent will want to implement the preferred consumption stream with the minimum amount of self-control, i.e. with the lowest discount factor δ, that will do the job. This lowest discount factor is strictly smaller in $\{a, c\}$ than in $\{a, b\}$. Thus, b, the "local competitor" to the chosen alternative a, is tempting in $\{a, b, c\}$, while the "globally largest temptation" c is not.

Exactly the opposite holds in the GP model in which menus are ranked according to the functional form

$$U(A) = \max_{x \in A} (u(x) + t(x)) - \max_{y \in A} t(y). \tag{14}$$

Here, consumption streams with larger present consumption are construed as *more* tempting rather than less; in particular, the preference pattern (12) implies that $t(c) > t(b)$. This in turn implies the ranking

$$\{a, b\} \succ \{a, b, c\} \sim \{a, c\}, \tag{15}$$

which is incompatible with (13). In particular, now c is tempting in $\{a, b, c\}$, but b is not.

This incompatibility is robust, and does not hinge on the linearity of the GP representation. In particular, it extends to the non-linear generalization (9) proposed by Fudenberg-Levine (2005) with strictly increasing k, where one has[21]

$$\{a, b\} \succ \{a, b, c\} \gtrsim \{a, c\}. \tag{16}$$

The contrast between the rankings (13) and (16) reveals a fundamental difference in the logic of the SOP model with convex feasible preferences and GP style models. To put it in a slogan, in the former "all temptation is local" while in the latter "all temptation is global".

Does this difference have a deeper meaning in terms of the nature of the implied *mechanism* of self-control? Since the GP model can be embedded fairly naturally in the SOP model as pointed out above, the difference cannot be attributed to a fundamental incompatibility of the two models as such. Instead, it is better accounted for *within* the SOP model as a difference between two types of second-order preferences. From this perspective, the inconsistency between (13) and (16) implies that GP style menu preferences are based on the formation of *non-convex* ex-post preferences. In the consumption-savings problem above, for example, this means that GP-style self-control cannot be

[21] The only point of contact between the two models is the Benhabib-Bisin (2004) model which is a limiting case of both, with k weakly but not strictly increasing in either representation. In that model, (12) implies that $\{a, b\} \sim \{a, b, c\} \sim \{a, c\}$.

understood as an ex-interim modification of agent's discount factors. Likewise, in a setting in which the alternatives are lotteries of final outcomes, it implies that a GP agent in this setting *cannot* have ex-post preferences of the expected-utility form.[22]

There is an element of paradox here. Why should ex-post preferences be systematically different *in character* from ex-ante preferences? While these observations do not disqualify GP-style menu preferences per se as a model of self-control, they seem to show that the associated ex-post preferences cannot be understood as coherent expressions of a coherent ex-post desire, hence that GP-style self-control cannot be understood as effortful modification of such desire (e.g. of the rate of time preference).[23]

If not as modification of desire, how can GP-style self-control understood then? To obtain a tentative answer, it is suggestive to consult the SOP rationalization of GP-style menu preferences offered in (8). It implies that a given alternative x can be optimally implemented in a menu A by a transformation of the default/temptation preference D that consists in lifting x just above the D-maximal alternative in A, leaving all other rankings as in D. Such a transformation can be viewed as the minimal change of D that enables the agent to choose x in A; by construction, it is highly specific to A, and is targeted to x, the intended choice, alone. By contrast, a modification of some underlying desire would in some systematic fashion affect the ranking of other alternatives as well, especially (under convexity) of locally competing ones. The targeted lifting of x is thus better interpreted as a *repression* of the "temptation" originating from the default preference rather than as the expression of a modified desire.

At an intuitive level, the distinction between these two modes of self-control seems quite appealing, and seems to track differences in real-word self-control problems. Self-control by repression may apply in cases in which the self-control problem originates in visceral impulses, in which the default desire demands gratification within a very short time horizon, and in which the ex post decision is cognitively simple. Choices among desserts at dinner would seem to fit this pattern, as would choices concerning the consumption of addictive goods, spontaneous shopping decisions, crimes of passion or opportunity, and surely many others.

[22] This statement is meaningful only *within* the SOP model, of course, in which ex-post preferences are endogeneous and depend on the menu. It is not contradicted by the fact that the revealed preference relation defined from the ex-post choice function C implied by GP menu preferences has the EU form. The latter is a derivative construct in the SOP model.

[23] For a different self-control account of the GP model with implied stochastic rather than deterministic ex-post choices, see Benaubou-Pycia (2002).

By contrast, self-control by desire modification should be more appropriate in situations in which the self-control problem originates in uneducated instincts or habits, in which the consequences of the ex-post decision are spread out over time, and in which the decision requires some planning or deliberation. Here, modification of desires through visualization of long-term consequences, deliberate weighing of pros and cons, detachment from emotion are plausible mechanisms of self-control. The life-time consumption-savings problem faced by most people would appear to fall in this category; other examples may be attempts to loosen or overcome the grip of decision traps such as loss aversion, regret avoidance, overconfidence, and wishful thinking; the transformation of dietary and fitness habits; etc. .

Obviously, these off-the-bat considerations are at best suggestive and deserve more careful elaboration. On the theoretical side, one obvious dimension requiring further attention is the dynamic one. In particular, the establishment of personal rules is a major mechanisms of self-control, and the distinction from and interaction with the repression and modification modes of self-control would need to be clarified. The dynamic dimension is all the more important inasmuch the modification (and sometimes also the repression) of desire will frequently have lasting effects, thus endowing self-control with the characteristics of an investment decision. On the empirical side, it would be important to relate these distinctions to extant distinctions in the psychological literature. Within the author's limited knowledge, the distinction between hot and cold states is perhaps the most closely relevant.

4. GENERAL CHARACTERIZATION OF THE SOP MODEL

SOP rationalizable menu preferences are characterized by the Upper Boundedness property introduced above which requires that for any menus $A, B \in \mathcal{M}$, the union $A \cup B$ is weakly inferior than one of the two. Upper Boundedness captures the principle that "unchosen alternatives can never help". Indeed, suppose the agent expects to choose from the menu $A \cup B$ an alternative x in A . Then he could have achieved the same outcome x in the menu A with the same ex-post preference P, hence with the same self-control effort; thus $A \succsim A \cup B$.

Thus Upper Boundedness is clearly necessary for SOP rationalizability. Less obviously, it is also sufficient.

Theorem 3 *On comprehensive $\mathcal{M}$, the menu-preference $\succsim$ is SOP rationalizable if and only if it satisfies Upper Boundedness.*

Upper Boundedness implies that no menu can be superior to the best alternative it contains, and thus excludes "preferences for flexibility" a la Kreps (1979). As pointed out by Dekel et al. (2005), beyond that Upper Boundedness also conflicts with "temptation uncertainty", which can be viewed as introducing a secondary preference for flexibility.[24] [25]

While a natural implication of self-control, Upper Boundedness is satisfied by many menu preferences that do not seem to naturally explicable in terms of self-control considerations. For example, a ranking of menus according to their *inverse* cardinality (that is: $A \succsim B$ iff $\#A \leq \#B$) is upper-bounded, but seems difficult to explain in terms of such considerations. Central to the notion of self-control is the opposition of ex-ante interests and ex-post temptations. From this perspective, it is hard to see how in some particular menu every alternative could be tempting, let alone how this could happen in every non-singleton menu as it does in the inverse cardinality ordering. In other words, it seems plausible to expect self-control-driven preferences to satisfy (at least) the following condition.

Condition 4 *For no $A \in \mathcal{M}$ it is the case that, for all $x \in A$, $A \prec A \backslash x$.*

If this is accepted, it means that the unstructured, bare-bones SOP model contains many second-order preference orderings that admit a self-control interpretation in at best a rather contrived sense. For example, the inverse-cardinality ranking described above is rationalized by second-order preferences represented by

$$V(x, P) = \#\{z : zPx\}.$$

Notice that this second-order preference ranks two extended outcomes (x, P) and (y, P) in strict opposition to there ranking by P itself: $(x, P) \trianglerighteq (y, P)$ if and only if yPx. While it is true that one can imagine certain diabolic or masochistic fantasies giving rise to such second-order preferences, their intuitive contradiction to any well-defined notions of self and self-interest makes it questionable whether such stories are meaningfully classified under the rubric of self-control.

In the following, we will therefore consider various "soundness" conditions on second-order preferences and show how they give rise to restrictions on menu preferences that entail Condition 4. In

[24] To see this, consider $X = \{a, b, c\}$, with commitment preferences $\{a\} \succ \{b\} \succ \{c\}$. Due to temptation uncertainty, it may easily happen that $\{a, b, c\} \succ \{a, c\} \succ \{b\}$, in violation of Upper Boundedness. This will happen if c is tempting a with low probability, and if the agent is able to resist this temptation by choosing b but not by choosing a. Adding b to $\{a, c\}$ is valuable as a stop-gap measure in case temptation strikes, even though $\{b\}$ is inferior to $\{a, c\}$ in isolation.

[25] See also Chatterjee and Krishna (2005) for a worked out model of menu-preferences driven by uncertainty about future temptations.

particular, we will show that the most generally applicable and compelling such condition, Limited Temptation, is characterized by a slight strengthening of Condition 4, thereby establishing a match between self-control-driven preferences over menus and well-structured second-order preferences over extended outcomes.

5. STRUCTURING THE SOP MODEL

In well-behaved second-order preferences, the evaluation of outcomes conditional on hypothetical ex-post preferences P (i.e. comparisons of the form (x, P) versus (y, P)) will be related to the content of those preferences. At one end of the spectrum, the conditional preferences "fully reflect" the ex-post preferences P.

Assumption 5 ***(Full Reflection)*** $(x, P) \trianglerighteq (y, P)$ *whenever* xPy.

Since Full Reflection rules out any conflict of interest between the ex-ante and ex-post selves, it entails an indirect utility ranking of menus, with commitment preferences given by $\{x\} \succsim \{y\}$ iff

$$\arg\max_{\trianglerighteq}\{(x, P) : P \in \mathcal{L}(X)\} \trianglerighteq \arg\max_{\trianglerighteq}\{(y, P) : P \in \mathcal{L}(X)\}.$$

While fully reflective preferences have no non-standard implications for preferences over menus as such, they may be interesting to study for their non-standard implications for preferences over the alternatives themselves. For example, in an intertemporal context, Becker and Mulligan (1997) model an agent who optimally chooses his discount rate, and whose second-order preferences satisfy Full Reflection.

The polar opposite of Full Reflection is the following assumption of Anti-Reflection.

Assumption 6 ***(Anti-Reflection, preliminary version)*** *There exists a weak order* W *such that* $(x, P) \trianglerighteq (y, P)$ *whenever* xWy, *and such that* $(x, P) \triangleright (y, P)$ *whenever* xWy *and not* yWx.

Anti-Reflection requires that, conditional on any P, outcomes are ranked according to a fixed "well-being" ordering W that is independent of the hypothetically formed ex-post preferences P. It captures the idea that the ex-ante self has a fixed, well-defined view of what benefits the agent overall; any modification of P has only the purpose of steering ex-post behavior in the right direction,

but does not influence the ex-ante perceived benefit. In view of this imperious attitude toward the ex-post self, we will call such second-order and menu preferences preferences "*with self-command*". An important example of second-order preferences with self-command are the additively separable ones of Section 3 characterized by a representation of the form $V(x, P) = u(x) - k(P)$.

It is easily verified that, under Anti-Reflection, the ordering W must coincide with the induced commitment preference $\succsim_1$.[26] Thus, to simplify the further exposition, one can replace W by $\succsim_1$ in the statement of Anti-Reflection. This leads to the following condition in which we additionally drop the strict part for technical reasons.[27]

Assumption 7 ***(Anti-Reflection)*** $(x, P) \trianglerighteq (y, P)$ *whenever* $\{x\} \succsim \{y\}$.

Intermediate between and generalizing these two polar cases is the following assumption of Partial Reflection. We will refer to the associated second-order and menu preferences as preferences "*with self-management*".

Assumption 8 ***(Partial Reflection)*** *There exists a weak order* W *on* X *such that*

$$(x, P) \trianglerighteq (y, P) \text{ whenever } xPy \text{ and } xWy. \tag{17}$$

In contrast to Full Reflection, Partial Reflection allows ex-post preferences to be overruled due to the existence of an ex-ante interest factor W that is ignored ex-post. And in contrast to Anti-Reflection, Partial Reflection allows first-order preferences P to influence the conditional valuation; that is, P is viewed by the ex-ante self not merely as causally determining which outcome is ultimately chosen, but also as contributing to the ex-ante desirability of that outcome. This will make sense in many cases: even if the ex-ante self views P as rationally deficient (infected by bias etc.), different ex-post preferences P will be accompanied by distinct psychological states of desire leading to distinct levels of desire-satisfaction. The actual enjoyment from a given culinary indulgence, for example, may be greatly reduced by accompanying worries about future weight-consequences. It is reasonable

[26] To see that under Anti-Reflection, $\{x\} \succsim \{y\}$ whenever xWy, take $(y, P) \in H(\{y\})$, $(x, Q) \in H(\{x\})$, and xWy. By Anti-Reflection, $(x, P) \trianglerighteq (y, P)$. Since also $(x, Q) \trianglerighteq (x, P)$, one obtains $(x, Q) \trianglerighteq (y, P)$, i.e. $\{x\} \succsim \{y\}$.

[27] The latter move makes no difference in the absence of non-trivial indifferences among singletons, a very common and not an unreasonable assumption in a finite setting.

To achieve a version of Theorem 10 for the stronger, "preliminary" version of Anti-Reflection, one would need to obtain a version of Theorem 17 in the Appendix that extends the asymmetric component of the given relation R_0; this does not appear to be straightforward.

for the ex-ante self to count this reduction of enjoyment as a real loss in well-being. In a related vein, Ainslie (2001) strongly emphasizes the "loss of appetite" as a major potential downside of the exercise of self-control.[28]

An interesting example of second-order preferences in the literature is Brunnermeier-Parker's (2005) theory of "optimal expectations" (see also Gollier 2005). Adapted to the present setting, it yields a theory of "optimal wishful thinking".[29] In a nutshell, in this theory ex-post preferences are determined by endogenously chosen subjective beliefs π. The ex-ante valuation of an act f $V(f,\pi)$ is made up of two components, the "anticipatory utility" $E_\pi u(f)$ (pertaining to the time interval preceding the resolution of the uncertainty) and the expected "realized utility" (pertaining to the time interval following the resolution of the uncertainty) $E_{\pi^*}u(f)$, the latter expectation being taken with respect to a fixed ex-ante probability π^*,

$$V(f,\pi) = E_\pi u(f) + \delta E_{\pi^*} u(f),$$

for some appropriate weighting factor $\delta > 0$. By contrast, ex-post preferences P use ex-post beliefs to evaluate both components; hence they rank acts according to $E_\pi u(f) + \delta E_\pi u(f)$, i.e. according to $E_\pi u(f)$. With W given by $E_{\pi^*}u(f)$, the second-order preferences associated with $V(f,\pi)$ satisfy Partial Reflection. Note that optimal-expectation based preferences are driven by a different type of trade-off than self-command preferences are: here, the cost of forming preferences in line with ex-ante beliefs is the opportunity cost of lost anticipatory utility, not the direct cost of self-control effort.

[28] As in the case of Anti-Reflection, it is of interest to eliminate the reference to the extraneous primitive W in the statement of the Partial Reflection assumption. This can be done as follows. Define a relation $W_{\trianglerighteq}$ on outcomes by setting $xW_{\trianglerighteq}y$ iff $(x,P) \rhd (y,P)$ and yPx; $W_{\trianglerighteq}$ summarizes all instances in which first-order preferences over outcomes are overriden ex-ante by conditional second-order preferences. Partial Reflection assumes that such overriding occurs in a consistent manner reflecting the existence of a consistent, well-defined iterest that is taken into account ex-ante but ignored ex post. Appealing to a finite version of Szpilrain's theorem, Partial Reflection as defined in the text is easily seen to be equivalent to acyclicity of $W_{\trianglerighteq}$. A conceptual advantage of this reformulation is the absence of any completeness requirement on $W_{\trianglerighteq}$; note that Full Reflection, for example, is equivalent to $W_{\trianglerighteq}$ being empty.

[29] Brunnermeier-Parker (2005) themselves interpret the choice of π as occurring behind the agent's back, for instance as made by "evolution".

By contrast, Epstein and Kopylov (2006) model an agent who anticipates temptation by wishful thinking (referred to by Epstein and Kopylov as minimization of 'cognitive dissonance'; however, menu preferences have a GP-like tempation representation of the form (9) (with linear ϕ) and have thus a natural SOP rationalization with self-command.

6. CHARACTERIZATIONS

Let us now turn to the restrictions on menu preferences implied by these axioms. Self-command preferences turn out to be characterized by the following axiom.

Axiom 9 ***(Singleton Monotonicity)*** *If* $\{x\} \succsim \{y\}$ *for all* $y \in A$, *then* $A \cup \{x\} \succsim A$.

Singleton Monotonicity is an intuitively transparent implication of self-command: clearly, if the agent chooses to switch to the ex-ante superior alternative x 'voluntarily' (at the original level of self-control) , that's beneficial ex-ante as well, if not, no harm is done. More pedantically, if menu preferences are SOP rationalizable, for an additional alternative x to hurt, x musts render the extended outcome(s) (y, P) chosen in A infeasible; this happens if and only if x is preferred to y under P. But in that case the extended outcome (x, P) is be feasible in $A \cup \{x\}$[30]. Since under Anti-Reflection this extended outcome is weakly preferred to the originally chosen one, $(x, P) \trianglerighteq (y, P)$. The addition of x does no harm after all, and $A \cup \{x\} \succsim A$, as asserted by Singleton Monotonicity.

Theorem 10 *On comprehensive* $\mathcal{M}$, $\succsim$ *is rationalizable by a second-order preference satisfying Anti-Reflection if and only if it satisfies Upper Boundedness and Singleton Monotonicity.*

In the presence of Upper Boundedness, Singleton Monotonicity is implied by Lower Boundedness. As the examples in section 3.3 and 3.4 have shown, it is much weaker.[31]

In contrast to self-command preferences, self-management preferences do not need to satisfy Singleton Monotonicity. The above derivation does not go through since the ex-ante self may now be made worse off by a switch to the alternative x; while superior in isolation (hence without the need for self-control), under Partial Reflection x need not be superior when self-control is exercised. In other words, the above argument breaks down since (x, P) may be second-order inferior to (y, P) (since y may well be ex-ante superior to x in terms of W).

The argument can be resurrected, however, if based on W rather than commitment preferences $\succsim_1$. Suppose $\trianglerighteq$ satisfies Partial Reflection with respect to W, and that xWy for all $y \in A$. Then if the agent prefers to switch from y to x at P, this switch is ex-ante beneficial, since by assumption x is also superior to y in terms of W; otherwise, no harm is done. Putting this slightly differently,

[30] Strictly speaking, in $Y_{A \cup \{x\}}$, of course.

[31] Note that Singleton Monotonicity weakens Lower Boundedness in two ways: first, the added menu is required to be a singleton, and second, this singleton needs to be weakly superior to any existing alternative as a singleton, not merely superior to the existing menu as a set ($A \precsim \{x\}$). (Upper Boundedness ensures that $A \precsim \{y\}$ for some $y \in A$).

if $\trianglerighteq$ satisfies Partial Reflection with respect to W, then a W-maximal alternative in A cannot be tempting in A or any of its subsets. Thus, leaving W unspecified, any A contains an alternative x such that x is tempting neither in A nor in any of its subsets. This is expressed by the following axiom of Limited Temptation.

Axiom 11 ***(Limited Temptation)*** *For all $A \in \mathcal{M}$, there exists $x \in A$ such that, for all $B \in \mathcal{M} : x \in B \subseteq A$, $B \backslash x \precsim B$.*

Limited Temptation in fact characterizes self-management preferences.

Theorem 12 *On comprehensive $\mathcal{M}$, $\succsim$ is rationalizable by a second-order preference satisfying Partial Reflection if and only if it satisfies Upper Boundedness and Limited Temptation.*

To illustrate the difference between self-command and self-control preferences over menus, consider the following example with 3 alternatives, $X = \{a, b, c\}$. Schematically, b is "safe"; c is disastrous and requires self-control to be avoided; a is irresistible vis-a-vis b and greatly rewarding in the absence of self-control efforts, but much less so in their presence. Menu-preferences and implied ex-post choices are as follows:

$$\{a\} \sim \{a, b^*\} \succ \{b\} \succsim \{b^*, c\} \succ \{a^*, b, c\} \sim \{a^*, c\} \succ \{c\}. \tag{18}$$

These menu preferences evidently satisfy Limited Temptation;[32] on the other hand, Singleton-Monotonicity is violated, since adding $\{a\}$ to the menu $\{b, c\}$ makes the agent worse off. This happens because while a is ex-ante superior to b when the agent can commit to either alternative and self-control efforts are therefore unnecessary, a is ex-ante inferior to b when c is feasible and self-control efforts are required to avoid it.

The preferences in (18) can be rationalized by second-order preferences with self-management as follows. There are two feasible preferences, D ("loose") and Q ("stern"), with the rankings $cDaDb$ and $aQbQc$. Extended outcomes are ranked as follows:

$$(a, D) \rhd (b, D) \rhd (b, Q) \rhd (a, Q) \rhd (c, D) \rhd (c, Q).$$

In choices among singletons, the agent will always adopt D; this implies in particular that $H(\{a\}) = (a, D) \rhd (b, D) = H(\{b\})$ and thus $\{a\} \succ \{b\}$. But in the presence of the disastrous alternative c, the agent needs to exercise self-control and switch to Q in order to avoid a choice of c ex-post.

[32] Simply note that b is not tempting in $\{a, b, c\}$.

Since b is better than a given Q, this implies $H(\{b,c\}) = (b,Q) \rhd H(\{a,b,c\}) = (a,Q)$ and thus $\{a,b,c\} \prec \{b,c\}$, in violation of Singleton Monotonicity.

The schematically described preference pattern can arise in a variety of contexts. As usual, a dieting story can be told. But other scenarios may be more interesting. Consider, for example, the decision problem of an agent on the verge of an extramarital affair. Self-control being a topic with strong Victorian overtones, it may be appropriate to link to the tradition of Madame Bovary, Anna Karenina and Effie Briest and refer to this agent as "wife" named Chloe; the presence of the third alternative a, however, gives the story a more contemporary feel. Chloe anticipates a meeting with the fancied (and presumed to be willing) "date", a meeting at which many things are possible. The impending affair may not materialize ("b"); it may become an fully-fledged, deeply involved relationship ("c"), with dire consequences for her marriage; finally, it could take the form of a (possibly exciting) short-term fling ("a") which would not endanger the marriage but would come at the price of some lies and bites of conscience. Chloe can now, alone, shape the plot with a cool head by two means: by precommitments that preclude certain outcomes to develop, and by deciding on the state of mind with which she enters the encounter: she could stay guarded, mindful of the priority of her marriage at any time; or she could fully open herself to passion. According to the stated preference relation, Chloe's ideal extended outcome would be a passionate fling (a,D). The reader is invited to fill in the missing details of the story him- or herself.

Judging from the models and examples that can be found in the literature, natural counterexamples to Limited Temptation which satisfy Upper Boundedness tend to be associated with counterexamples to the following more elementary condition of Weak Lower Boundedness which says any menu is weakly preferred to the worst alternative it contains.

Axiom 13 ***(Weak Lower Boundedness)*** *For all $A \in \mathcal{M}$, there exists $x \in A$ such that $A \succsim \{x\}$.*

Since Weak Lower Boundedness seems compelling a priori as a necessary condition property of menu-preferences driven by self-control considerations, this observation further strengthens the intuitive appeal of Limited Temptation and tends to support an identification of menu preferences driven by self-control considerations with menu preferences preferences "with self-management" as defined here.

An interesting example of menu preferences violating Weak Lower Boundedness is Dekel et al.s (2005) multiple temptations model with the representation

$$u(x,A) = u(x) + \sum_i (v_i(x) - \max_{y \in A} v_i(y)).$$

Weak Lower Boundedness will be violated in this model whenever there exist alternatives x and y and temptations i and j such that $v_i(x) > v_i(y)$, $v_j(x) < v_j(y)$, and $u(x) + \sum_i v_i(x) = u(y) + \sum_i v_i(y)$, since then $\{x\} \sim \{y\} \succ \{x,y\}$. In the lottery framework underlying the model, such pairs of lotteries fail to exist only for menu preferences with rather special structure.[33] It follows that, in contrast to the GP model, multiple temptation preferences typically cannot be rationalized by second-order preferences with self-management. This may not be very surprising since there is an important disanalogy in the notion of temptation in GP and DLR: while the single temptation utility in GP can be interpreted in the SOP model as the default choice disposition that would govern ex-post choices in the absence of self-control efforts, no such interpretation is available for the multiple simultaneous temptations in the DLR model.[34]

[33] Indeed, the menu preferences specified by Dekel et al. to explain the example of section 3.3 above violate Lower Boundedness.

[34] See also Sarver (2005) who proposes a reinterpretation of the multiple temptations model in terms of anticipated regret.

APPENDIX: PROOFS

A1. Background: Indirect Utility on General Domains

We will begin with a result on indirect utility orderings on general, unstructured domains. Let Z a domain of "elements", and $\mathcal{S} \subseteq 2^Z \backslash \varnothing$ a domain of sets, with $\geq$ a weak order on $\mathcal{S}$. The result will be applied to the case of $Z = X \times \mathcal{L}(X)$ and $\mathcal{S} = \{Y_A : A \in \mathcal{M}\}$, with $Y_A := \{(P(A), P) : P \in \mathcal{L}(X)\}$. To ease into it, we will begin with a special case.

Axiom 14 *If $\cup B_i \supseteq A$, then, for some i, $B_i \geq A$.*

Proposition 15 *$\geq$ satisfies Axiom 14 if and only if there exists a weak order R on Z such that*

$$A \geq B \text{ iff } (\arg\max{}_R A)\, R\, (\arg\max{}_R B).$$

Proof. Special case of Theorem 17 below.

The result to be shown strengthens Axiom 14 to guarantee that the weak order R on Z extends a given partial order (transitive and reflexive relation) R_0. Define $D(S, R_0) := \{x \in Z |$ for some $y \in S : yR_0x\}$.

Axiom 16 *If $D(\cup B_i, R_0) \supseteq A$, that is: if, for all $x \in X$ there exists $y \in \cup B_i$ such that yR_0x, then , for some i, $B_i \geq A$.*

Theorem 17 *$\geq$ satisfies Axiom 16 if and only if there exists a weak order $R \supseteq R_0$ on Z such that*

$$A \geq B \text{ iff } (\arg\max{}_R A)\, R\, (\arg\max{}_R B).$$

Proof of Theorem 17.

Necessity is obvious. For sufficiency, let

$$Y_{\mathcal{M}} := \{x \in Z | \text{for no } B, A \in \mathcal{M} : A > B \text{ and } x \in D(B, R_0)\}. \tag{19}$$

The construction of R is based on the following Lemma.

Lemma 18 *$S \geq T$ for all $T \in \mathcal{M}$ if and only if $Y_{\mathcal{M}} \cap S \neq \varnothing$.*

Proof of Lemma. Evidently, if S is not $\geq -$top ranked then $S \cap Y_{\mathcal{M}} = \varnothing$ (simply set $B = S$, and choose $A > S$; such A exists by the completeness of $\geq$).

Conversely, suppose that $S \cap Y_{\mathcal{M}} = \varnothing$. Then, by the definition of $Y_{\mathcal{M}}$, for each $x \in S$, there exist menus B_x and A_x in $\mathcal{M}$ such that i) $x \in D(B_x, R_0)$, and ii) $A_x > B_x$. By Axiom 16 and i), for some $x \in S$, $S \leq B_x$. By ii), therefore also $S < A_x$, showing that S is not $\geq -$top ranked. $\square$

Construct the ranking R as follows. Define inductively a nested sequence $\{\mathcal{M}_k\}$ in $\mathcal{M}$ by setting $\mathcal{M}_1 := \mathcal{M}$ and $\mathcal{M}_k := \{S \in \mathcal{M}_{k-1} : S \cap Y_{\mathcal{M}_{k-1}} = \varnothing\}$. Note that, by construction, the $\{Y_{\mathcal{M}_k}\}$ form a partition of Z. For $x \in Y_{\mathcal{M}_k}$ and $y \in Y_{\mathcal{M}_{k'}}$, define

$$xRy \text{ if and only if } k \leq k'.$$

Let $\geq_R$ denote theIU ordering induced by R; we claim that $\geq = \geq_R$ and $R \supseteq R_0$.

For the first claim, S is $\geq -$top ranked if and only $S \cap Y_{\mathcal{M}} \neq \varnothing$ by Lemma 18 which holds if and only if S is $\geq_R -$top ranked by the construction of R. Thus $\geq = \geq_R$ by a straightforward inductive argument.

Clearly, for any x, y such that yR_0x, by the transitivity of R_0, for any k, $x \in Y_{\mathcal{M}_k}$ implies $y \in Y_{\mathcal{M}_k}$, hence $x \in Y_{\mathcal{M}_k}$ and $y \in Y_{\mathcal{M}_{k'}}$ imply $k' \leq k$, establishing yRx as desired. ■

A2. A Master-Result for Menu-Preferences

We now apply Theorem 17 to sets of extended outcomes. Say that a partial order $\trianglerighteq_0$ on $X \times \mathcal{L}(X)$ is *outcome-based* if $(x, P) \trianglerighteq_0 (y, Q)$ implies $P = Q$, for all x, y, P, Q.

Given a menu preference $\succsim$, define the induced preference $\succsim_{\mathcal{Y}}$ over "menus of extended outcomes" $Y \in \mathcal{Y}$ by

$$Y \succsim_{\mathcal{Y}} Y' \text{ whenever there exist } A, B \in \mathcal{M} \text{ such that } Y = Y_A \text{ and } Y' = Y_B \text{ and } A \succsim B;$$

note that the A and B referred to in this definition are unique.

Axiom 19 *Let $\{B_i\}_{i \in I}$ be an arbitrary family of menus and $A \in \mathcal{M}$. If, for all $P \in \mathcal{L}(X)$, there exists i such that $(P(B_i), P) \trianglerighteq_0 (P(A), P)$, then $B_i \succsim A$ for some $i \in I$.*

Theorem 20 ***(Master Theorem)*** *Suppose that the partial order $\trianglerighteq_0$ is outcome-based. Then $\succsim$ has a SOP rationalization such that $\trianglerighteq \supseteq \trianglerighteq_0$ if and only if it satisfies Axiom 19.*

Proof of Theorem 20.

From the definition of D,

$$D(\cup Y_{B_i}, \trianglerighteq_0) = \{(x,P) \in X \times \mathcal{L}(X) | \text{for some } (y,Q) \in \cup Y_{B_i} : (y,Q) \trianglerighteq_0 (x,P)\};$$

by the outcome-basedness of $\trianglerighteq_0$, one can assume that $Q = P$ in the r.h.s. . It follows that $D(\cup Y_{B_i}, \trianglerighteq_0) \supseteq Y_A$ if and only if for all $P \in \mathcal{L}(X)$, there exists i such that $(P(B_i), P) \trianglerighteq_0 (P(A), P)$. Hence $\succsim_{\mathcal{Y}}$ satisfies Axiom 16 if and only if $\succsim$ satisfies Axiom 19. Theorem 20 thus follows immediately from Theorem 17. ■

Theorem 20 provides a schematic master result that will now be used obtain the three main results of the paper by plugging in three different specifications of $\trianglerighteq_0$ and simplifying the corresponding version of Axiom 19.

A3. Existence of a General SOP Representation

We first derive a version of Theorem 3 for arbitrary domains of menus, taking $\trianglerighteq = \trianglerighteq_\varnothing$, where $\trianglerighteq_\varnothing$ is the "vacuous" (reflexive) relation given by

$$(x,P) \trianglerighteq_\varnothing (y,Q) \text{ if and only if } x = y \text{ and } P = Q.$$

Axiom 21 *Let $\{B_i\}_{i \in I}$ be an arbitrary family of menus and $A \in \mathcal{M}$. If $A = \cup_i B_i$, then $B_i \succsim A$ for some $i \in I$.*

Theorem 22 *On arbitrary $\mathcal{M}$, $\succsim$ on $\mathcal{M}$ has a SOP representation if and only if satisfies Axiom 21.*

Proof of Theorem 22. In view of the definition of $\trianglerighteq_\varnothing$, the result is an immediate consequence of the Master Theorem (Theorem 20) and the following Lemma.

Lemma 23 *The following two statements are equivalent:*

i) For all $P \in \mathcal{L}(X)$, $P(A) \in \{P(B_i)\}_{i \in I}$;

ii) $A = \cup_i B_i$.

Proof of Lemma.

ii) implies i). Immediate from fact that P linear order.

i) implies ii). Suppose that ii) is false. Then there exists $x \in A$ such that, for all $B_i \ni x$, $B_i \backslash A \neq \varnothing$. Let $P \in \mathcal{L}(X)$ be such that $P(A) = x$ and such that yPz whenever $y \in A^c$ and $z \in A$. Since by

assumption $B_i \backslash A \neq \varnothing$ whenever $x \in B_i$, then by the construction of P one has $P(B_i) \in A^c \subseteq \{x\}^c$; on the other hand, if $x \notin B_i$, trivially $P(B_i) \neq x$. It follows that $\{P(B_i)\}_{i \in I} \subseteq \{x\}^c$. □■

Proof of Theorem 3.

Theorem 3 is an immediate consequence of the following Lemma.

Lemma 24 *If $\mathcal{M}$ is comprehensive, Axiom 21 is equivalent to Upper Boundedness.*

Proof. Upper Boundedness is evidently a special case of Axiom 21. To see that it in fact implies Axiom 21 if $\mathcal{M}$ is comprehensive, take any family $\{B_i\}$ and A such that $A = \cup_i B_i$. Assume w.l.o.g. that $B_1 \succsim B_2 \succsim ... \succsim B_n$. By comprehensiveness of $\mathcal{M}$, the sets $C_j := \bigcup_{i \leq j} B_i$ are contained in $\mathcal{M}$. By Upper Boundedness, $B_1 = C_1 \succsim C_2 \succsim ... \succsim C_n = A$. □■

A4. Second-Order Preferences with Self-Command

Let W be a partial order on X, and define the outcome-based partial order $\trianglerighteq_W$ on $X \times \mathcal{L}(X)$ by

$$(x, P) \trianglerighteq_W (y, Q) \text{ if and only if } xWy \text{ and } P = Q.$$

Say that B **α-covers x for A** if $x \in B$ and, for all $y \in B \backslash A$, yWx; B **β-covers x for A** if $x \notin B$ and, for all $y \in B$, yWx; and B **covers x for A** if B α-covers x for A or B β-covers x for A.

Axiom 25 *If for all $x \in A$ there exists i such that B_i covers x for A, then for some i, $B_i \succsim A$.*

Axiom 26 ***(W-Addition)*** *If xWy for all $y \in A$, then $A \cup \{x\} \succsim A$.*

Axiom 27 ***(W-Substitution)*** *if yWx for all $y \in A$, then $\{x\} \precsim A$.*

Theorem 28 *i) $\succsim$ on arbitrary $\mathcal{M}$ has a SOP representation with $\trianglerighteq \supseteq \trianglerighteq_W$ if and only if satisfies Axiom 25.*

ii) If W is a weak order, $\succsim$ on comprehensive $\mathcal{M}$ has a SOP representation with $\trianglerighteq \supseteq \trianglerighteq_W$ if and only if it satisfies Upper Boundedness, W-Addition and W-Substitution.

Proof of Theorem 28, Part i).

The result is an immediate consequence of the Master Theorem (Theorem 20) combined with Lemma 29 below.

Lemma 29 *The following two statements are equivalent for arbitrary A and $\{B_i\}$:*

i) For all $P \in \mathcal{L}(X)$, there exists $i \in I$ such that $P(B_i)\ W\ P(A)$;

ii) For all $x \in A$, there exists $i \in I$ such that B_i covers x for A.

Proof of Lemma.

ii) implies i).

Take any $P \in \mathcal{L}(X)$ and B_i such that B_i covers $P(A)$ for A.

Case i): $P(A) \in B_i$. Then, by choice consistency, $P(B_i) \in P(A) \cup (B_i \backslash A)$ whence $P(B_i)\ W\ P(A)$ since B_i α-covers x.

Case ii): $P(A) \notin B_i$. Then $P(B_i)\ W\ P(A)$ since B_i β-covers x.

i) implies ii).

Proof via modus tollens. Fix some $x \in A$ that is not covered by any B_i. Let $J = \{i \in I : B_i \ni x\}$.

Thus, by definition, for any $i \in J$, there exists $z_i \in B_i \backslash A$ such that not $z_i W x$ (*),

and, for any $i \in I \backslash J$, there exists $z_i \in B_i$ such that not $z_i W x$ (**).

Let Z be a selection of such z_i, $Z := \{z_i : i \in I\}$. Let P be a linear ordering that ranks alternatives as follows: First, the alternatives in the set $Z \cap A^c$, in arbitrary order. Second, x. Third, the alternatives in the set $Z \cap A$. Fourth, all other alternatives.

By construction, for any $z \in Z$ and $y \in X$: if yPz then $y \in Z \cup \{x\}$; hence, for all $i \in I$, $P(B_i) \in Z \cup \{x\}$.

We claim that for all $i \in I$, $P(B_i) \neq x$, hence $P(B_i) \in Z$. Indeed, for i such that $x \in B_i$, $z_i \in B_i \backslash A$ by (*), whence $z_i P x$ by the definition of P, and thus $P(B_i) \neq x$. On the other hand, for i such that $x \notin B_i$, trivially if $P(B_i) \neq x$. Since therefore $P(B_i) \in Z$ for all $i \in I$, for no $i \in I$, $P(B_i)\ W\ x$ by (*) and (**). Since also $x = P(A)$ by the construction of P, for no $i \in I$, $P(B_i)\ W\ P(A)$, as desired. □

Proof of Theorem 28, Part ii).

This follows immediately from Part i) and Lemma 30 below.

Lemma 30 *If $\mathcal{M}$ is comprehensive and W is a weak order $\succsim$ satisfies Axioms 25 if and only if it satisfies Upper Boundedness, W-Addition and W-Substitution.*

Proof of Lemma.

Necessity is straightforward. For sufficiency, suppose that, for all $x \in A$, there exists i such that B_i W-covers x for A. We need to show that, for some $i \in I$, $B_i \succsim A$.

For $i \in I$, let $C_i = \{x \in A \cap B_i : yWx \text{ for all } y \in B_i \backslash A\}$, and let $\mathcal{C}_1 = \{C_i\}_{i \in I}$. Also, let $\mathcal{C}_2$ denote the family of singletons $\{x\}$ for $x \in A$ such that, there exists $i \in I$ such that yWx for all $y \in B_i$.

By the hypothesis on the family $\{B_i\}$, evidently $\bigcup (\mathcal{C}_1 \cup \mathcal{C}_2) = A$. By Upper Boundedness, comprehensiveness of $\mathcal{M}$ and Lemma 24, therefore $C \succsim A$ for some $C \in \mathcal{C}_1 \cup \mathcal{C}_2$.

If $C = C_i \in \mathcal{C}_1$, in view of Lemma 31,i) just below and since W is a weak order, by W-Addition $B_i \succsim C$, hence $B_i \succsim A$ by transitivity.

On the other hand, if $C = \{x\} \in \mathcal{C}_2$, by the definition of $\mathcal{C}_2$ there exists $i \in I$, such that yWx for all $y \in B_i$. In view of Lemma 31,ii) just below and since W is a weak order, by W-Substitution $B_i \succsim C$, hence $B_i \succsim A$ by transitivity. $\square$

Lemma 31 *Suppose $\mathcal{M}$ is comprehensive and W is a weak order. Then*

i) $\succsim$ satisfies W-Addition if and only if $B \succsim A$ whenever, for all $x \in A$, B α-covers x for A;

ii) $\succsim$ satisfies W-Substitution if and only if $B \succsim A$ whenever, for all $x \in A$, B β-covers x for A.

The straightforward proof is left to the reader. ■

Proof of Theorem 10.

Theorem 10 follows immediately from substituting the commitment ordering $\succsim_1$ for W in part ii) of Theorem 28 together with the observation that Singleton Monotonicity (i.e. $\succsim_1$-Addition) implies $\succsim_1$-Substitution, which, for weak orders $\succsim$, is easily seen to be equivalent to Weak Lower Boundedness. ■

A5. Second-Order Preferences with Self-Management

Again, let W be a partial order on X, and define the outcome-based partial order $\trianglerighteq_{WP}$ on $X \times \mathcal{L}(X)$ by

$$(x, P) \trianglerighteq_{WP} (y, Q) \text{ if and only if } xWy,\ xPy \text{ and } P = Q.$$

Axiom 32 *If, for all $x \in A$, there exists i such that B_i α-covers x for A, then for some i, $B_i \succsim A$.*

Theorem 33 *i) $\succsim$ on arbitrary $\mathcal{M}$ has a SOP representation with $\trianglerighteq \supseteq \trianglerighteq_{WP}$ if and only if satisfies Axiom 32.*

ii) If W is a weak order, $\succsim$ on comprehensive $\mathcal{M}$ has a SOP representation with $\trianglerighteq \supseteq \trianglerighteq_{WP}$ if and only if satisfies Upper Boundedness and W-Addition.

Proof of Theorem 33, Part i). The result is an immediate consequence of Theorem 20 combined with the following Lemma.

Lemma 34 *The following two statements are equivalent for arbitrary A and $\{B_i\}$:*

i) For all $P \in \mathcal{L}(X)$, there exists $i \in I$ such that $P(B_i)$ W $P(A)$ and $P(B_i)$ P $P(A)$;

ii) For all $x \in A$, there exists $i \in I$ such that B_i α-covers x for A.

Proof of Lemma.

ii) implies i).

Take any $P \in \mathcal{L}(X)$ and B_i such that B_i α-covers $P(A)$ for A. Since $P(A) \in B_i$, $P(B_i)$ P $P(A)$. Moreover, by choice consistency, $P(B_i) \subseteq P(A) \cup (B_i \backslash A)$ whence $P(B_i)$ W $P(A)$ as well by α-coverage.

i) implies ii). Proof via modus tollens.

Fix some $x \in A$ that is not α-covered by any B_i.

Let $J := \{i \in I : B_i \ni x\}$. By assumption, for any $i \in J$, there exists $z_i \in B_i \backslash A$ such that not $z_i W x$. Let Z be a selection of such z_i, $Z := \{z_i : i \in J\}$.

Let P be an ordering that ranks alternatives as follows: First, the alternatives in the set Z, in arbitrary order. Second, x. Third, all other alternatives, in arbitrary order. Evidently, since $Z \subseteq A^c$, $x = P(A)$.

By construction, for any $i \in I$ such that $B_i \cap Z \neq \varnothing$, $P(B_i) \in Z$ by choice consistency, and therefore not $P(B_i)$ W $P(A)$. On the other hand, for any $i \in I$ such that $B_i \cap Z = \varnothing$, $x \notin B_i$ by the definition of P; hence from the construction of P, not $P(B_i)$ P $P(A)$.

Thus for no $i \in I$, $P(B_i)$ $W \cap P$ $P(A)$. $\square$

Proof of Theorem 33, Part ii). ii) follows immediately from Part i) and the following Lemma.

Lemma 35 *If $\mathcal{M}$ is comprehensive and W is a weak order, $\succsim$ satisfies Axioms 32 if and only if it satisfies Upper Boundedness and W-Addition.*

Proof of Lemma.

Necessity is straightforward. For sufficiency, suppose that, for all $x \in A$, there exists $i \in I$ such that B_i α-covers x for A. We need to show that, for some i, $B_i \succsim A$.

For $i \in I$, let $C_i = \{x \in A \cap B_i : yWx \text{ for all } y \in B_i \backslash A\}$. By the hypothesis on the family $\{B_i\}$, $\cup C_i = A$. By Upper Boundedness, comprehensiveness of $\mathcal{M}$ and Lemma 24, therefore $C_i \succsim A$ for some $i \in I$. In view of Lemma 31, i), by the definition of C_i, W-Addition and the fact that W is a weak order, $B_i \succsim C_i$, hence $B_i \succsim A$ by transitivity. □■

The proof of Theorem 12 relies on the following general result on "acyclic indirect utility representations". Let $\mathcal{S}$ denote a family of non-empty subsets of some finite set Z, and $>$ be a relation on $\mathcal{S}$.

Theorem 36 *The following two statements are equivalent:*

i) There exists a linear order P such that

$$A > B \text{ implies } (\arg\max{}_P A)\, P\, (\arg\max{}_P B)$$

ii) If $A_i > B_i$ for all i, then $\bigcup A_i \backslash \bigcup B_i \neq \varnothing$.

Proof of Theorem 36.

The implication from i) to ii) is straightforward.

The converse is proved by induction on the cardinality of Z. The claim is trivial for $\#Z = 1$; suppose that it is satisfied for $\#Z \leq n-1$. Consider $>$ on Z with $\#Z = n$. From ii), it is evident that there is $z^* \in Z$ such that there do not exist $A, B \in 2^Z \backslash \{\varnothing\}$ with $z^* \in B$ and $A > B$.

Consider now the preference relation $>^*$ on $Z^* := Z \backslash \{z^*\}$ defined as the restriction of $>$ to $Z \backslash \{z^*\}$, i.e. , for any $A, B \in 2^{Z \backslash \{z^*\}} \backslash \{\varnothing\}$,

$$A >^* B \text{ if and only if } A > B.$$

Evidently, $>^*$ satisfies ii). By induction assumption, there exists a linear order P^* on $Z \backslash \{z^*\}$ such that, for any $A, B \in 2^{Z \backslash \{z^*\}} \backslash \{\varnothing\}$,

$$A > B \text{ implies } (\arg\max{}_{P^*} A)\, P^*\, (\arg\max{}_{P^*} B)\,.$$

Let P be the linear order on Z uniquely defined by putting z^* on top, and equal to P^* on $Z \backslash \{z^*\}$. Evidently, $A > B$ implies $B \in 2^{Z \backslash \{z^*\}} \backslash \{\varnothing\}$ by definition of z^*. Since also trivially $A \in 2^{Z \backslash \{z^*\}} \backslash \{\varnothing\}$ or $A \ni z^*$, we therefore have, for any $A, B \in 2^Z \backslash \{\varnothing\}$,

$$A > B \text{ implies } (\arg\max{}_P A)\, P\, (\arg\max{}_P B)\,,$$

as desired. □

Proof of Theorem 12.

The necessity of Limited Temptation has been shown in the text.

Define $A > \{x\}$ iff $A \cup \{x\} \prec A$. By Limited Temptation, for any family $\{A_i\}$ and $\{x_i\}$ such that $A_i > \{x_i\}$ for all $i \in I$, $\bigcup_i A_i \backslash \bigcup_i \{x_i\} \neq \varnothing$. For evidently, if to the contrary $\bigcup A_i \subseteq \{x_i\}_{i \in I}$, then Limited Temptation would fail at the menu $\{x_i\}_{i \in I}$. Hence by Theorem 36 above, there exists a linear order W such that, for any $A \in \mathcal{M}$, and $x \in X$, $A \cup \{x\} \prec A$ implies that there exists $y \in A$ such that $yW\dot{x}$. Arguing by modus tollens, this means that $\succsim$ satisfies $W-$Addition. Hence the existence of a partially reflective second-order preference $\trianglerighteq$ rationalizing $\succsim$ follows from Theorem 28. ■

REFERENCES

[1] Ainslie, G. (2001): "Breakdown of the Will", Cambridge University Press.

[2] Becker, G. and C.B. Mulligan (1997): "The endogenous determination of time preference," *Quarterly Journal of Economics* 112, 729—758.

[3] Benabou, R. and M. Pycia (2002): "Dynamic Inconsistency and Self-Control," *Economics Letters* 77, 419-424.

[4] Benabou, R. and J. Tirole (2004): "Willpower and Personal Rules," *Journal of Political Economy*, 112, 848-887.

[5] Benhabib, J. and A. Bisin (2005): "Modelling Internal Commitment-Mechanisms and Self-Control: A Neuroeconomics Approach to Consumption-Savings Decisions", *Games and Economic Behavior* 52, 460-492.

[6] Bernheim, B.D. and A. Rangel (2004): "Addiction and Cue-Triggered Decision Processes," mimeo, forthcoming in the *American Economic Review*.

[7] Brunnermeier, M. and J. Parker (2005): "Optimal Expectations", *American Economic Review*, 95(4), 1092-1118.

[8] Epstein, L. and I. Kopylov (2005): "Cognitive Dissonance and Choice", University of Rochester Working Paper.

[9] Chatterjee, K. and V. Krishna (2005): "Menu Choice, Environmental Cues and Temptation: A 'Dual Self" Approach to Self-Control', Penn State University Working Paper.

[10] Dekel, E., B. Lipman and A. Rusticini (2005): "Temptation-Driven Preferences", Boston University Working Paper.

[11] Fudenberg, D. and D. Levine (2005): "A Dual Self Model of Impulse Control", Harvard University Working Paper, forthcoming in the *American Economic Review*.

[12] Gollier, C. (2004): "Optimal Positive Thinking and Decisions under Risk", mimeo Toulouse.

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

[14] Gul, F. and W. Pesendorfer (2006): "The Simple Theory of Temptation," mimeo Princeton.

[15] Kopylov, I. (2005): "Temptations in General Settings", mimeo.

[16] Kreps, D. (1979): "A Representation Theorem for 'Preference for Flexibility'," Econometrica 47, 565-577.

[17] Krusell, P, Kuruscu, B. and T. Smith (2005): "Temptation and Taxation," mimeo Princeton.

[18] Loewenstein, G. and O'Donoghue, E. (2005): "Animal Spirits: Affective and Deliberative Processes in Economic Behavior", mimeo.

[19] Machina, M. (1984) "Temporal Risk and the Nature of Induced Preferences," *Journal of Economic Theory* 33 199-231.

[20] Mas-Colell, A. , Whinston, M. and J. Green (1995), *Microeconomic Theory*, Oxford University Press.

[21] Miao, J. (2005) "Option Exercise with Temptation," mimeo, Boston University.

[22] Muraven, M, and R. Baumeister (2000) "Self-Regulation and Depletion of Limited Resources: Does Self-Control Resemble a Muscle?" *Psychological Bulletin*, 126, 27-259.

[23] Muraven, M, D. Tice, and R. Baumeister (1998) "Self-Control as a Limited Resource," *Journal of Personality and Social Psychology*, 74, 774-789.

[24] Noor, J. (2005) "Temptation, Welfare and Revealed Preferences," mimeo, Rochester.

[25] Noor, J. (2006): "Menu-Dependent Self-Control", Boston University Working Paper.

[26] O'Donoghue, T. and M. Rabin (1999): "Doing it Now or Doing it Later," *American Economic Review* 89, 103-124.

[27] O'Donoghue, T. and M. Rabin (2001): "Choice and Procrastination," *Quarterly Journal of Economics,* 121-160.

[28] Ozdenoren, E., S. Salant and D. Dan Silverman (2006): "Willpower and the Optimal Control of Visceral Urges," Institute for Advanced Study working paper #69.

[29] Sarver, T. (2005): "Anticipating Regret: Why Fewer Options May Be Better", mimeo.

[30] Sen, A. (1971): "Choice Functions and Revealed Preference", *Review of Economic Studies* 38, 307-317.

[31] Shiv, B. and A. Fedorikhin (1999): "Heart and Mind in Conflict: The Interplay of Affect and Cognition in Consumer Decision Making", *Journal of Consumer Research* 26, 72-89.

[32] Strotz, R. H. (1955): "Myopia and inconsistency in dynamic utility maximization", *Review of Economic Studies*, 23, 165-180.

[33] Thaler, R. and H.M. Shefrin (1981): "An Economic Theory of Self-Control," *Journal of Political Economy*, 89, 392-406.

[34] Ward, A. and T. Mann (2000): "Don't Mind If I Do: Disinhibited Eating Under Cognitive Load" *Journal of Personality and Social Psychology*, 753-763.