Temptation–Driven Preferences[1]

Eddie Dekel[2] Barton L. Lipman[3] Aldo Rustichini[4]

February 2005
Current Draft

[1]We thank Massimo Marinacci, Ben Polak, Phil Reny and numerous seminar audiences for helpful discussions. We also thank the NSF for financial support for this research.

[2]Economics Dept., Northwestern University, and School of Economics, Tel Aviv University E–mail: dekel@nwu.edu.

[3]Boston University. E–mail: blipman@bu.edu. This work was begun while this author was at the University of Wisconsin.

[4]University of Minnesota. E–mail: arust@econ.umn.edu.

Abstract

"My own behavior baffles me. For I find myself not doing what I really want to do but doing what I really loathe." Saint Paul

What behavior can be explained using the hypothesis that the agent faces temptation but is otherwise a "standard rational agent"? In earlier work, Gul–Pesendorfer [2001] use a set betweenness axiom to restrict the set of preferences considered by Dekel, Lipman, and Rustichini [2001] to those explainable via temptation. We argue that set betweenness rules out plausible and interesting forms of temptation. We propose a pair of alternative axioms called DFC, *desire for commitment*, and AIT, *approximate improvements are tempting*. DFC characterizes temptation as situations where given any set of alternatives, the agent prefers committing herself to some particular item from the set rather than leaving herself the flexibility of choosing later. AIT says that if adding an option to a menu improves the menu, it is because that option is chosen under some circumstances, not because its presence reduces the extent to which other options are tempting. We show that these axioms characterize a natural generalization of the Gul–Pesendorfer representation.

1 Introduction

What potentially observable behavior would indicate that the decision maker faces temptation? Phrased differently, what behavior can we explain using the hypothesis that the agent faces temptation but is otherwise a "standard rational agent"? We use the phrase *temptation–driven* to refer to behavior explainable in this fashion.

By "temptation," we mean that the agent has some view of what is normatively correct, what she *should* do, but has other, conflicting desires which must be reconciled with the normative view in some fashion. We view this conflict as being among the available options, but where the extent to which any one option is tempting the agent is independent of what other options are available. Also, while there is undoubtedly an element of arbitrariness in this modeling choice, we allow the possibility that the extent or nature of temptation is random, but do not allow similar randomness regarding what is normatively preferred. We take this approach to separate temptation–driven behavior from issues of flexibility.

Our approach builds on earlier work by Gul–Pesendorfer [2001] (henceforth GP) and Dekel, Lipman, and Rustichini [2001] (DLR). DLR consider a large set of preferences (behavior) which includes preferences driven by a desire for flexibility, preferences driven by temptation, and preferences combining both considerations. GP focus on temptation alone by considering a particular subset of these preferences which are naturally explained using only temptation. We argue that the axiom they use to identify temptation, set betweenness, rules out plausible and interesting forms of temptation. In this sense, they identify a subset of the temptation–driven preferences.

We propose a pair of alternative axioms called DFC, *desire for commitment*, and AIT, *approximate improvements are tempting*. We show that these axioms characterize a natural generalization of the GP representation. DFC simply says that given any set of alternatives, the agent at least weakly prefers to commit herself to some option from this set rather than retaining the flexibility to choose from the set later. In this sense, DFC is exactly the statement that there is no value to flexibility but the agent may fear being tempted to choose "inappropriately." AIT says that if adding an option to a menu improves the menu, it is because the option added is chosen under some conditions, not because its presence reduces the extent to which other options are tempting. Given the interpretation of the axioms and the intuitive nature of the representation they generate, we conclude that DFC and AIT yield a natural way to identify from the large set considered by DLR those preferences which are temptation–driven.

We also give some special cases of the main representation and the additional axioms which correspond to these. As we explain, these special cases have a natural interpretation as restrictions on the kinds of temptation faced by the agent.

In the next section, we present the basic model and state our research goals more precisely. In the process, we sketch the relevant results in DLR and GP. In Section 3, we give examples to motivate the issues and illustrate the kinds of representations we are interested in. In Section 4, we give representation results. Because DFC is a simpler axiom than AIT and because it is a convenient step in the analysis, we also state the representation generated by adding only this axiom to the original DLR axioms. Section 5 contains characterizations of some special cases. In Section 6, we briefly discuss directions for further research.

2 The Model

Let B be a finite set of *prizes* and let $\Delta(B)$ denote the set of probability distributions on B. A typical subset of $\Delta(B)$ will be referred to as a *menu* and denoted x (or $\tilde{x}$, x', $\bar{x}$, y, etc.), while a typical element of $\Delta(B)$, a *lottery*, will be denoted by β. The agent has a preference relation $\succ$ on the set of closed nonempty subsets of $\Delta(B)$ which is denoted X. Given menus x and y and a number $\lambda \in [0,1]$, let

$$\lambda x + (1-\lambda)y = \{\beta \in \Delta(B) \mid \beta = \lambda\beta' + (1-\lambda)\beta'', \text{ for some } \beta' \in x, \beta'' \in y\}$$

where, as usual, $\lambda\beta' + (1-\lambda)\beta''$ is the probability distribution over B giving b probability $\lambda\beta'(b) + (1-\lambda)\beta''(b)$.

The relevant axioms used in DLR [2001] are:

Axiom 1 (Weak Order) *$\succ$ is asymmetric and negatively transitive.*

Axiom 2 (Continuity) *The strict upper and lower contour sets, $\{x' \subseteq \Delta(B) \mid x' \succ x\}$ and $\{x' \subseteq \Delta(B) \mid x \succ x'\}$, are open (in the Hausdorff topology).*

Axiom 3 (Independence) *If $x \succ x'$, then for all $\lambda \in (0,1]$ and all $\bar{x}$,*

$$\lambda x + (1-\lambda)\bar{x} \succ \lambda x' + (1-\lambda)\bar{x}.$$

We refer the reader to DLR for further discussion of these axioms.

We use a finite state version of one of the representations in DLR. More specifically, we use

Definition 1 *A **finite additive EU representation** is a finite set S, a state–dependent utility function $U : \Delta(B) \times S \to \mathbf{R}$, and a measure μ with full support on S such that (i) $V(x)$ defined by*

$$V(x) = \sum_{s \in S} \mu(s) \max_{\beta \in x} U(\beta, s)$$

is continuous and represents $\succ$ and (ii) each $U(\cdot, s)$ is an expected–utility function in the sense that

$$U(\beta, s) = \sum_{b \in B} \beta(b) U(b, s).$$

It is important to emphasize that the measure μ need not be a probability measure. In particular, we can have $\mu(s) < 0$ for some s. The meaning and importance of this point is discussed in more detail below.

One can show that adding a "finiteness" axiom to the original DLR axioms characterizes the finite additive EU representation. While not particularly elegant, one version of such an axiom is

Axiom 4 (Finiteness) *There exists a finite N such that for every menu x, there is a subset $x' \subseteq x$ with N elements or fewer such that $x' \sim x$.*

It is not difficult to use the approach of DLR to show[1]

Theorem 1 *The preference $\succ$ has a finite additive EU representation if and only if it satisfies weak order, continuity, independence, and finiteness.*

While the form above is the one used in DLR (except for the finiteness), it will prove convenient to rewrite the representation as follows. Let $\{s_1, \ldots, s_I\}$ denote the set of *positive states* — those with $\mu(s) > 0$ — and let $\{s_{I+1}, \ldots, s_{I+J}\}$ denote the set of *negative states* — those with $\mu(s) < 0$. For $i = 1, \ldots, I$, let $w_i(\beta) = U(\beta, s_i)\mu(s_i)$. For $j = 1, \ldots, J$, let $v_j(\beta) = U(\beta, s_{I+j})|\mu(s_{I+j})|$. Then we can write

$$V(x) = \sum_{i=1}^{I} \max_{\beta \in x} w_i(\beta) - \sum_{j=1}^{J} \max_{\beta \in x} v_j(\beta).$$

[1] Aside from the addition of the finiteness axiom, this result differs from that in DLR [2001] in two respects. First, DLR included a requirement that S be nonempty as part of the definition of an additive EU representation and correspondingly included a nontriviality axiom. Second, DLR required that no state be redundant implying, in particular, that there is no s such that $U(\cdot, s)$ is a constant function.

If there were no negative states, this representation would have an obvious interpretation. Think of the w_i's as different utility functions the agent might have at some later date when she will choose from the menu she picks today. At the point when she will make this choice, she will know which of the w_i's is her utility function and, naturally, will choose the item from the menu which maximizes this utility. Her *ex ante* evaluation of the menu, then, is simply the expected value of the maximum. If each of the w_i's is equally likely, we obtain the value above. This is exactly the interpretation originally offered by Kreps [1979, 1992] who first considered preferences over sets as a way of modeling preference for flexibility. Obviously, though, the presence of the negative states makes this interpretation awkward.

One way to reach a clearer understanding of this representation, then, is to rule out the negative states. DLR show that the axiom which does this is one of Kreps' main axioms, namely monotonicity.

Axiom 5 (Monotonicity) *If $x \subset x'$, then $x' \succeq x$.*

A simple modification of DLR's proof shows

Theorem 2 *The preference $\succ$ has a finite additive EU representation with a positive measure μ if and only if it satisfies weak order, continuity, independence, finiteness, and monotonicity.*

Rephrased, if $\succ$ satisfies monotonicity, then the additive EU representation contains no negative states. This result is the generalization of Kreps' [1979] representation theorem.

Intuitively, monotonicity is the statement that the agent always at least weakly values flexibility. Such an agent either is not concerned about temptation or, at least, values flexibility so highly as to outweigh such considerations. In this case, the finite additive EU representation is easy to understand as describing a forward–looking agent with some beliefs about what her possible future needs are.

GP take a different approach. They first recognized that temptation and self–control could be studied using this sets of lotteries framework if one does not impose monotonicity. If the agent might be tempted in the future to consume something she currently doesn't want herself to consume, this is revealed by a preference for commitment, not flexibility. GP's [2001] representation theorem differs from Theorem 1 in that they add an axiom which they call *set betweenness*[2]:

[2]Their result differs in two other ways as well. First, set betweenness turns out to imply the finiteness

Axiom 6 (Set Betweenness) *If $x \succeq y$, then $x \succeq x \cup y \succeq y$.*

To understand this axiom, suppose the agent is deciding where to eat lunch and wishes to consume a healthy meal. Think of x, y, and $x \cup y$ as the menus available at the three possible restaurants. Suppose x consists only of a single healthy food item, say broccoli, while y consists only of some fattening food item, say french fries. Then the fact that the agent wants to consume a healthy meal suggests $x \succ y$. How should the agent rank the menu $x \cup y$ relative to the other two? A natural hypothesis is that the third restaurant would fall in between the other two in the agent's ranking. It would be better than the menu with only french fries since the agent might choose broccoli given the option. On the other hand, the third menu would be worse than the menu with only broccoli since the agent might succumb to temptation or, even if she didn't succumb, might suffer from the costs of maintaining self–control in the face of the temptation. Hence $x \succ x \cup y \succ y$, in line with what set betweenness requires.

The relevant representation in GP is

Definition 2 *A* self control representation *is a pair of functions (u, v), $u : \Delta(B) \to \mathbf{R}$, $v : \Delta(B) \to \mathbf{R}$, such that each is an expected utility function and the function V_{GP} defined by*

$$V_{GP}(x) = \max_{\beta \in x}[u(\beta) + v(\beta)] - \max_{\beta \in x} v(\beta)$$

represents $\succ$.

GP show

Theorem 3 *A self control representation exists if and only if the preference satisfies weak order, continuity, independence, and set betweenness.*

To interpret this representation, first note that we can think of u as the "commitment preference" — that is, what the agent would choose if she could commit herself *ex ante*. Specifically, $V(\{\beta\}) = u(\beta)$ for any β. In light of this, following GP, we interpret u as describing the agent's view of what is normatively appropriate.[3] For any menu x and any $\beta \in x$, let

$$c(\beta, x) = \left[\max_{\beta' \in x} v(\beta')\right] - v(\beta).$$

axiom, so they do not impose it separately. This claim is implied by their Lemma 2, page 1422. Second, they consider a more general setting in that they assume B is compact, not finite.

[3]See, however, Noor [2004] for a cogent critique of this interpretation and an alternative formulation.

Intuitively, c is the foregone utility according to v from choosing β from x instead of the β' which is optimal according to v. It is easy to see that

$$V_{GP}(x) = \max_{\beta \in x}[u(\beta) - c(\beta, x)].$$

In this form, it is natural to interpret c as the cost of the self–control needed to choose β from x. Given this, v is naturally interpreted as the temptation utility since it is what determines the self–control cost.

Another way to understand the GP representation is to relate it to DLR. It is easy to see that the self–control representation is exactly a finite additive EU representation with one positive state and one negative state where $w_1 = u + v$ and $v_1 = v$. In this sense, the only difference between GP's representation and the DLR representation with one positive and one negative state is a change of variables.

One way to think about these results is to begin by considering the set of preferences satisfying weak order, continuity, independence, and finiteness. For brevity, we will refer to these as *DLR preferences*. Intuitively, if we consider the subset of these DLR preferences which are monotonic, we are restricting attention to agents who value flexibility but are not affected by temptation. It seems very natural to call such preferences *flexibility–driven*, both because the axiom and the representation it generates seem to describe such an agent. Putting the point differently, flexibility–driven agents are those whose behavior can be explained by flexibility considerations alone.

Analogously, we will refer to preferences which are explained by a concern about temptation but no value to flexibility *per se* as *temptation–driven*. It seems natural to say that the subset of DLR preferences that satisfy set betweenness are temptation–driven preferences. However, set betweenness does not appear to be as obvious a statement of "temptation–driven preferences" as monotonicity is for "flexibility–driven." In fact, it is not hard to give examples of behavior which appears to be temptation–driven but where set betweenness is violated. This suggests that set betweenness is stronger than just a restriction to temptation–driven preferences. Our goal in this paper is to identify and give a representation theorem for the full class of temptation–driven DLR preferences.

3 Motivating Examples and Some Alternative Representations

In this section, we give two examples to illustrate our argument that set betweenness is stronger than a restriction to temptation–driven preferences. We also use these examples to suggest some representations that may be of interest.

Example 1.

Consider an agent who is trying to diet and so would like to commit herself to eating only broccoli. There are two kinds of snacks available: chocolate cake and high fat potato chips. Let b denote the broccoli, c the chocolate cake, and p the potato chips. The following ranking seems quite natural:

$$\{b\} \succ \{b, c\}, \{b, p\} \succ \{b, c, p\}.$$

That is, the agent would like to commit herself to eating only broccoli, so $\{b\}$ is the best of these four menus. If she has both broccoli and a fattening snack available, the temptation of the snack will lower her utility, so $\{b, c\}$ and $\{b, p\}$ are both worse than $\{b\}$. If she has broccoli and *both* fattening snacks available, she is still worse off since two snacks are harder to resist than one.

Two snacks could be worse than one for at least two reasons. First, it could be that the agent is unsure what kind of temptation will strike. If the agent would be in a mood for a salty snack, then she may be able to control herself easily if only the chocolate cake is available as an alternative to broccoli. Similarly, if she is in the mood for a sweet snack, she may be able to control herself if only the potato chips are available. But if she has both available, she is more likely to be hit by a temptation she cannot avoid. Second, even if she resists temptation, the psychological cost of self–control seems likely to be higher in the presence of two snacks than in the presence of one.[4]

This preference violates set betweenness. Note that $\{b, c, p\}$ is strictly worse than $\{b, c\}$ and $\{b, p\}$ even though it is the union of these two sets. Hence set betweenness implies that two temptations can *never* be worse than each of the temptations separately. In GP, temptation is one dimensional in the sense that any menu has a most tempting option and only this option is relevant to the self–control costs.

It is not hard to give generalizations of GP's representation that can model either of the two reasons stated above for two snacks to be worse than one. To see this, define utility functions u, v_1, and v_2 by

	u	v_1	v_2
b	3	2	2
c	0	0	6
p	0	6	0

Define V_1 by the following natural generalization of GP:

$$V_1(x) = \frac{1}{2} \sum_{i=1}^{2} \left[\max_{\beta \in x} [u(\beta) + v_i(\beta)] - \max_{\beta \in x} v_i(\beta) \right].$$

[4] GP [2001, 1408–1409] mention this possibility as one reason why set betweenness may be violated.

Equivalently, let

$$c_i(\beta, x) = \left[\max_{\beta' \in x} v_i(\beta')\right] - v_i(\beta).$$

Then

$$V_1(x) = \frac{1}{2}\sum_{i=1}^{2} \max_{\beta \in x}[u(\beta) - c_i(\beta, x)].$$

Intuitively, the agent doesn't know whether the temptation that will strike is the one described by v_1 and cost function c_1 (where she is most tempted by the potato chips) or v_2 and cost function c_2 (where she is most tempted by the chocolate cake) and gives probability 1/2 to each possibility. It is easy to verify that this gives $V_1(\{b\}) = 3$, $V_1(\{b, c\}) = V_1(\{b, p\}) = 3/2$, and $V_1(\{b, c, p\}) = 0$, yielding the ordering suggested above.

Alternatively, define V_2 by a different generalization of GP:

$$V_2(x) = \max_{\beta \in x}[u(\beta) + v_1(\beta) + v_2(\beta)] - \max_{\beta \in x} v_1(\beta) - \max_{\beta \in x} v_2(\beta).$$

Here we can think of cost of choosing β from menu x as

$$c(\beta, x) = \left[\max_{\beta \in x} v_1(\beta) + \max_{\beta \in x} v_2(\beta)\right] - v_1(\beta) - v_2(\beta).$$

It is not hard to see that this cost function has the property that resisting two temptations is harder than resisting either separately. More specifically, it is easy to verify that $V_2(\{b\}) = 3$, $V_2(\{b, c\}) = V_2(\{b, p\}) = -1$, and $V(\{b, c, p\}) = -5$, again yielding the ordering suggested above.

Example 2.

Consider again the dieting agent facing multiple temptations, but now suppose the two snacks available are high fat chocolate ice cream (i) and low fat chocolate frozen yogurt (y). In this case, it seems natural that the agent might have the following rankings:

$$\{b, y\} \succ \{y\} \quad \text{and} \quad \{b, i, y\} \succ \{b, i\}.$$

In other words, the agent would rather have a chance of sticking to her diet rather than committing herself to violating it so $\{b, y\} \succ \{y\}$. Also, if the temptation of the ice cream is unavoidable, it's better to also have the low fat frozen yogurt around. If so, then when temptation strikes, the agent may be able to resolve her hunger for chocolate in a less fattening way.

Again, GP cannot have this. This is not a violation of set betweenness but instead a violation of the combination of set betweenness and independence. To see why this cannot occur in their model, note that

$$V(\{b, y\}) = \max\{u(b) + v(b), u(y) + v(y)\} - \max\{v(b), v(y)\}$$

while $V(\{y\}) = u(y) = u(y) + v(y) - v(y)$. Obviously, $\max\{v(b), v(y)\} \geq v(y)$. So $V(\{b, y\}) > V(\{y\})$ requires $\max\{u(b) + v(b), u(y) + v(y)\} > u(y) + v(y)$ or $u(b) + v(b) > u(y) + v(y)$. Given this,

$$\max\{u(b) + v(b), u(i) + v(i), u(y) + v(y)\} = \max\{u(b) + v(b), u(i) + v(i)\}.$$

Since

$$\max\{v(b), v(i), v(y)\} \geq \max\{v(b), v(i)\},$$

we get $V(\{b, i, y\}) \leq V(\{b, i\})$. That is, we must have $\{b, i\} \succeq \{b, i, y\}$.[5]

Intuitively, in GP, $\{b, y\} \succ \{y\}$ implies that the agent will never choose frozen yogurt when broccoli is available. Hence the only effect frozen yogurt can have when broccoli is available is to increase self–control costs. The possibility that y could be a compromise against some worse temptation is not allowed when $\{b, y\} \succ \{y\}$.

For a simple generalization of GP which allows the intuitive preference suggested above, define

	u	v
b	6	0
i	0	8
y	4	6

and let

$$V_3(x) = \frac{1}{2} \max_{\beta \in x} u(\beta) + \frac{1}{2} \left\{ \max_{\beta \in x} [u(\beta) + v(\beta)] - \max_{\beta \in x} v(\beta) \right\}.$$

Intuitively, there is a probability of 1/2 that the agent avoids temptation and chooses according to the commitment preference u. With probability 1/2, the agent is tempted, however, and has a preference of the form characterized by GP. This gives $V_3(\{b, y\}) = 5 > 4 = V_3(\{y\})$ and $V_3(\{b, i, y\}) = 5 > 3 = V_3(\{b, i\})$, in line with the intuitive story.

The three representations used in these examples share certain features in common. First, all are finite additive EU representations. That is, all the preferences involved satisfy weak order, continuity, independence, and finiteness. While we do not wish to argue that these axioms are innocuous, it is not obvious that temptation should require some violation of these properties (though see Section 6). Second, in all cases, the representation is written in terms of the utility functions for the negative states and u, the commitment utility. Equivalently, we can write the representation in terms of the commitment utility and various possible cost functions where these costs are generated from different possible temptations.

[5] We cannot show this directly from the axioms. We do know, however, that it cannot be demonstrated from set betweenness alone — independence is essential to this conclusion. More specifically, this preference is consistent with set betweenness if independence is violated or independence if set betweenness is violated.

Intuitively, the different negative states from the additive EU representation identify the different temptations. The various positive states then correspond to different ways these temptations might combine to affect the agent. However, all the positive states share a common view of what is "normatively best" as embodied in u. In this sense, there is no uncertainty about "true preferences" and hence no "true" value to flexibility, only uncertainty about temptation.

A general kind of representation with these properties is

Definition 3 *A* temptation representation *is a function V_T representing $\succ$ such that*

$$V_T(x) = \sum_{i=1}^{I} q_i \max_{\beta \in x} [u(\beta) - c_i(\beta, x)]$$

where $q_i > 0$ for all i, $\sum_i q_i = 1$, and

$$c_i(\beta, x) = \left[\sum_{j \in J_i} \max_{\beta' \in x} v_j(\beta') \right] - \sum_{j \in J_i} v_j(\beta)$$

where u and each v_j is an expected–utility function.

Note that $\sum_i q_i = 1$ implies that $V_T(\{\beta\}) = u(\beta)$, so u is the commitment utility.

Intuitively, we can think of each c_i as a cost of self–control, describing one way the agent might be affected by temptation. In this interpretation, q_i gives the probability that temptation takes the form described by c_i.

We can think of this as generalizing GP in two directions. First, more than one temptation can affect the agent at a time. That is, the cost of self–control may depend on more than one temptation utility. Second, the agent is uncertain which temptation or temptations will affect her.

A less interpretable representation which is useful as an intermediate step is

Definition 4 *A* weak temptation representation *is a function V_w representing $\succ$ such that*

$$V_w(x) = \sum_{i=1}^{I'} q_i \max_{\beta \in x} [u(\beta) - c_i(\beta, x)] + \sum_{i=I'+1}^{I} \max_{\beta \in x} [-c_i(\beta, x)]$$

where $q_i > 0$ for all i, $\sum_i q_i = 1$, and

$$c_i(\beta, x) = \left[\sum_{j \in J_i} \max_{\beta' \in x} v_j(\beta') \right] - \sum_{j \in J_i} v_j(\beta)$$

where u and each v_j is an expected–utility function.

Obviously, a temptation representation is a special case of a weak temptation representation where $I' = I$.

As we will see, the weak temptation representation makes a natural midway point between the temptation representation and the finite additive EU representation. On the other hand, it lacks the natural interpretation of the temptation representation.

4 New Axioms and Results

An axiom which seems to be a natural part of a definition of temptation–driven is

Axiom 7 (DFC: Desire for Commitment) *A preference $\succ$ satisfies DFC if for every x, there is some $\alpha \in x$ such that $\{\alpha\} \succeq x$.*

Intuitively, this axiom seems to be a necessary condition to say that a preference is temptation–driven. The axiom says that there is no value to flexibility associated with x, only potential costs due to temptation leading the agent to choose some point other than α.

On the other hand, this axiom only says that flexibility is not valued. It does not say anything about when commitment is valued. The second axiom identifies a key circumstance in which commitment is strictly valuable.

To motivate the second axiom, consider what it means for commitment to be strictly valuable. By DFC, there is an $\alpha \in x$ such that $\{\alpha\} \succeq x$. Commitment is strictly valuable if there is an $\alpha \in x$ with $\{\alpha\} \succ x$. Intuitively, this says that the agent believes that there is some situation in which either she does not succeed in choosing α (or something better for her diet) from x or else she does choose α but finds it costly to do so. To identify when commitment is strictly valuable, then, requires us to identify a circumstance where this is true. For brevity, whenever there is an $\alpha \in x$ such that $\{\alpha\} \succ x$, we say that x *is tempted*.

In light of this, consider the following situation. Suppose we have a menu x with the property that adding β to x strictly improves the menu for the agent. That is, $x \cup \{\beta\} \succ x$. In such a case, we say β is *an improvement for* x. When β is an improvement for x, it must be true that there is some situation in which β is chosen from the menu $x \cup \{\beta\}$. If it is never chosen, then the only effect it has is to increase the cost of self–control needed to resist β. In particular, by hypothesis, the presence of β cannot reduce the pull of other temptations. Given this, though, suppose there is $\alpha \in x$ such that $\{\alpha\} \succ \{\beta\}$. That is,

β is chosen from $x \cup \{\beta\}$ in some situation even though there is something else on the menu better for the agent's diet. In this case, it seems clear that $x \cup \{\beta\}$ is tempted.

Similarly, consider any $y \subset x$ which contains some α such that $\{\alpha\} \succ \{\beta\}$. Since β is chosen from $x \cup \{\beta\}$ in some situation, surely it is also chosen from $y \cup \{\beta\}$ in this same situation. Hence $y \cup \{\beta\}$ must also be tempted.

As an aside, we remark that this argument relies on a kind of independence of irrelevant alternatives. That is, we are arguing that if β is chosen from a set in some situation, then it is chosen from any subset containing it in that same situation. As Noor [2004] suggests by example, this is not necessarily an appropriate assumption for modeling temptation.

The axiom we need is slightly stronger. In addition to applying to any β which is an improvement for x, it applies to any β which is an approximate improvement for x. Because of this, we call the axiom AIT, *approximate improvements are tempting*. Formally, define β to be an *approximate improvement for* x if

$$\beta \in \mathrm{cl}\left(\{\beta' \mid x \cup \{\beta'\} \succ x\}\right)$$

where cl denotes closure. Then we have

Axiom 8 (AIT: Approximate Improvements are Tempting) *If β is an approximate improvement for x, $y \subseteq x$, and $\alpha \in y$ satisfies $\{\alpha\} \succ \{\beta\}$, then we have $y \cup \{\beta\}$ is tempted.*

Theorem 4 *$\succ$ has a temptation representation if and only if it has a finite additive EU representation and satisfies DFC and AIT.*

As mentioned earlier, the weak temptation representation, while not as interpretable as the temptation representation, is a natural intermediate point between the finite additive EU representation and the temptation representation. More specifically, in the course of proving Theorem 4, we also show

Theorem 5 *$\succ$ has a weak temptation representation if and only if it has a finite additive EU representation and satisfies DFC.*

5 Special Cases

In this section, we characterize the preferences corresponding to two special cases of temptation representations, specifically two of the three representations used in the examples. These special cases are of interest in part because of the way the required conditions relate to GP's set betweenness axiom. Also, these special cases can be thought of as narrowing the "allowed" forms of temptation in easily interpretable ways.

First, consider a representation of the form

$$V_{NU}(x) = \max_{\beta \in x} \left[u(\beta) + \sum_{j=1}^{J} v_j(\beta) \right] - \sum_{j=1}^{J} \max_{\beta \in x} v_j(\beta)$$

which we call a *no–uncertainty representation*. Equivalently,

$$V_{NU}(x) = \max_{\beta \in x} [u(\beta) - c(\beta, x)]$$

where

$$c(\beta, x) = \left[\sum_{j=1}^{J} \max_{\beta' \in x} v_j(\beta') \right] - \sum_{j=1}^{J} v_j(\beta).$$

Note that this representation differs from the general temptation representation by assuming that $I = 1$ — that is, that the agent knows exactly which temptations will affect her. Hence we call this a no–uncertainty representation to emphasize the idea that it is what we obtain when there is no uncertainty about the temptations. This representation, then, generalizes GP only by allowing the agent to be affected by multiple temptations.

If the preference has a finite additive EU representation with one positive state, then we can rewrite it in the form of a no–uncertainty representation by a generalization of the change of variables discussed in Section 2. Specifically, suppose we have a representation of the form

$$V(x) = \max_{\beta \in x} w_1(\beta) - \sum_{j=1}^{J} \max_{\beta \in x} v_j(\beta).$$

The commitment utility u is defined by $u(\beta) = V(\{\beta\}) = w_1(\beta) - \sum_j v_j(\beta)$. Hence we can change variables to rewrite V in the form of V_{NU}.

The no–uncertainty representation turns out to correspond to a particular half of set betweenness. Specifically,

Axiom 9 (Positive Set Betweenness) $\succ$ *satisfies positive set betweenness if whenever* $x \succeq y$, *we have* $x \succeq x \cup y$.

For future use, we define the other half similarly:

Axiom 10 (Negative Set Betweenness) $\succ$ *satisfies negative set betweenness if whenever* $x \succeq y$*, we have* $x \cup y \succeq y$*.*

We have

Lemma 1 *Suppose* $\succ$ *has a finite additive EU representation. Then it has such a representation with one positive state if and only if it satisfies positive set betweenness.*

To see the intuition, suppose $\succ$ satisfies positive set betweenness and suppose $x \succeq y$. Then $x \cup y$ is bounded "on the positive side" in the sense that $x \succeq x \cup y$. Hence the flexibility of being able to choose between x and y has only negative consequences. That is, the flexibility to choose between x and y cannot be better than x, though it can, conceivably, be worse than y. Hence the uncertainty the agent faces regarding her tastes is entirely on the negative side. This implies that there may be multiple negative states but can only be one positive one. The multiple negative states correspond to the different temptations the agent faces; the fact that there is only one positive state means that these affect him in exactly one way, so there is no uncertainty.

Using the change of variables discussed above, it is easy to see that this lemma yields

Theorem 6 $\succ$ *has a no–uncertainty representation if and only if it has a finite state additive EU representation and satisfies positive set betweenness.*

One can modify the proof of Lemma 1 in obvious ways to show

Lemma 2 *Suppose* $\succ$ *has a finite state additive EU representation. Then it has such a representation with one negative state if and only if it satisfies negative set betweenness.*

Theorem 3 is obviously a corollary to Lemmas 1 and 2.

A second special case takes Lemma 2 as its starting point. This representation has one negative state but many positive states which differ only in the strength of temptation in that state. Specifically, we define an *uncertain strength of temptation representation* to be one which takes the form

$$V_{US}(x) = \sum_{i=1}^{I} q_i \left[\max_{\beta \in x} [u(\beta) + \gamma_i v(\beta)] - \gamma_i \max_{\beta \in x} v(\beta) \right]$$

where $q_i > 0$ for all i and $\sum_i q_i = 1$. Equivalently,

$$V_{US}(x) = \sum_i q_i \max_{\beta \in x} [u(\beta) - \gamma_i c(\beta, x)]$$

where

$$c(\beta, x) = [\max_{\beta' \in x} v(\beta')] - v(\beta).$$

In this representation, the temptation is always v, but the strength of the temptation (as measured by γ_i) is random. The probability that the strength of the temptation is γ_i is given by q_i. In a sense, this representation allows uncertainty but to the minimum possible extent. Intuitively, the one negative state implies that there is only one temptation. Different positive states correspond to different extents to which the agent is affected by this temptation.

We have

Theorem 7 $\succ$ *has an uncertain strength of temptation representation if and only if it has a finite state additive EU representation and satisfies DFC and negative set betweenness.*

6 Conclusion

There are several interesting issues left to explore. In the previous section, we gave two specializations of the general representation to more specific assumptions on the nature of temptation. Naturally, there are numerous other possible directions of interest along similar lines.

Another direction of interest is the extent to which the temptation representation is identified and the preference equivalents of changes in the representation. For example, it is easy to use results in DLR to show that one preference has a representation with a larger set of negative states than another if and only if it values commitment more in a certain sense. Since temptation representations have more structure than additive EU representations, there may be new kinds of comparisons of interest.

Finally, our characterization of behavior which can be explained by temptation is carried out within the set of DLR preferences. While weak order and finiteness seem unrelated to issues of temptation, the assumptions of continuity and independence arguably eliminate some temptation–related behavior. Hence it may be useful to consider weaker forms of these axioms.

Regarding continuity, GP show that at least one common model of temptation requires continuity to be violated. To be specific, suppose the agent evaluates a menu x according

to

$$\max_{\beta \in B_v(x)} u(\beta)$$

where $B_v(x)$ is the set of v maximizers in x. Intuitively, the agent expects his choice from the menu to be determined by his later self with utility function v, where his later self breaks ties in favor of the current self. As GP demonstrate, in the absence of very specific relationships between u and v, such a representation cannot satisfy continuity.

Regarding independence, there are several temptation–related issues which may lead to violations of this axiom. To give one example, guilt may lead the agent to prefer randomization, a phenomenon inconsistent with independence. To see the point, consider a dieter in a restaurant faced with a choice between a healthy dish and a tempting, unhealthy dish. Such a dieter may strictly prefer adding to the menu a randomization between the two. With such an option available, the dieter can choose the lottery and have some chance of consuming the unhealthy dish with less guilt than if it had been chosen directly.[6]

[6]We thank Phil Reny for suggesting this example. The example has a strong resemblance to the "Machina's mom" story in Machina [1989].

A Notational Conventions

Throughout the Appendix, we write u, v_j, etc., to denote the vector giving the payoffs to the pure outcomes associated with utility function u, v_j, etc. We will always write these as column vectors. Because there are n pure outcomes, then, these are n by 1. We will write lotteries as 1 by n row vectors, so $\beta \cdot u = u(\beta)$, etc. Also, $\mathbf{1}$ denotes the n by 1 vector of 1's.

B Proof of Theorem 5

The following lemma is critical.[7]

Lemma 3 *Suppose $\succ$ has a finite additive EU representation of the form*

$$V(x) = \sum_{i=1}^{I} \max_{\beta \in x} w_i(\beta) - \sum_{j=1}^{J} \max_{\beta \in x} v_j(\beta).$$

Define u by $u(\beta) = V(\{\beta\})$, so $u = \sum_i w_i - \sum_j v_j$. Suppose $\succ$ satisfies DFC. Then there are positive scalars a_i, $i = 1, \ldots, I$, and b_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$ and scalars c_i, $i = 1, \ldots, I$ such that $\sum_i a_i = \sum_i b_{ij} = 1$ for all j and

$$w_i = a_i u + \sum_j b_{ij} v_j + c_i \mathbf{1}$$

for all i.

Proof. Suppose not. Let Z denote the set of nI by 1 vectors $(z_1', \ldots, z_I')'$ such that

$$z_i = a_i u + \sum_j b_{ij} v_j + c_i \mathbf{1}, \quad \forall i$$

for scalars a_i, b_{ij}, and c_i satisfying the conditions of the lemma. So if the lemma does not hold, the vector $(w_1', \ldots, w_I')' \notin Z$. Since Z is obviously closed and convex, the separating hyperplane theorem implies that there is a vector p such that

$$p \cdot \begin{pmatrix} w_1 \\ \vdots \\ w_I \end{pmatrix} > p \cdot \begin{pmatrix} z_1 \\ \vdots \\ z_I \end{pmatrix}, \quad \forall \begin{pmatrix} z_1 \\ \vdots \\ z_I \end{pmatrix} \in Z.$$

[7] This result can be seen as a generalization of the Harsanyi aggregation theorem (Harsanyi [1955]). See Weymark [1991] for an introduction to this literature.

Write $p = (p_1, \ldots, p_I)$ where each p_i is a 1 by n vector. So

$$\sum_i p_i \cdot w_i > \sum_i p_i \cdot z_i, \quad \forall \begin{pmatrix} z_1 \\ \vdots \\ z_I \end{pmatrix} \in Z.$$

Equivalently,

$$\sum_i p_i \cdot w_i > \sum_i a_i p_i \cdot u + \sum_j \sum_i b_{ij} p_i \cdot v_j + \sum_i c_i p_i \cdot \mathbf{1}$$

for any a_i, b_{ij}, and c_i such that $a_i \geq 0$ for all i, $b_{ij} \geq 0$ for all i and j, and $\sum_i a_i = \sum_i b_{ij} = 1$ for all j. Since c_i is arbitrary in both sign and magnitude, we must have $p_i \cdot \mathbf{1} = 0$ for all i. If not, we could find a c_i which would violate the inequality above.

Also, for every choice of $a_i \geq 0$ such that $\sum_i a_i = 1$,

$$\max_i p_i \cdot u \geq \sum_i a_i p_i \cdot u$$

with equality for an appropriately chosen $(a_1, \ldots, a_I)$. Similarly, for any nonnegative b_{ij}'s with $\sum_i b_{ij} = 1$,

$$\max_i p_i \cdot v_j \geq \sum_i b_{ij} p_i \cdot v_j$$

with equality for an appropriately chosen $(b_{1j}, \ldots, b_{Ij})$. Hence the inequality above implies

$$\sum_i p_i \cdot w_i > \max_i p_i \cdot u + \sum_j \max_i p_i \cdot v_j.$$

Write p_i as $(p_{1i}, \ldots, p_{ni})$. Without loss of generality, we can assume that $|p_{ki}| \leq 1/n$ for all k and i. (Otherwise we could divide both sides of the inequality above by $n \max_{k,i} |p_{ki}|$ and redefine p_i to have this property.) Let β denote the probability distribution $(1/n, \ldots, 1/n)$. For each i, let $\alpha_i = p_i + \beta$. Note that $\alpha_{ki} = p_{ki} + 1/n$ and so $\alpha_{ki} \geq 0$ for all k, i. Also, $\alpha_i \cdot \mathbf{1} = p_i \cdot \mathbf{1} + \beta \cdot \mathbf{1} = 1$. Hence each α_i is a probability distribution. Substituting $\alpha_i - \beta$ for p_i,

$$\sum_i \alpha_i \cdot w_i - \sum_i \beta \cdot w_i > \max_i \alpha_i \cdot u - \beta \cdot u + \sum_j \max_i \alpha_i \cdot v_j - \sum_j \beta \cdot v_j.$$

By definition of u, $\sum_i w_i = u + \sum_j v_j$. Hence this is

$$\sum_i \alpha_i \cdot w_i - \sum_j \max_i \alpha_i \cdot v_j > \max_i \alpha_i \cdot u.$$

Let $x = \{\alpha_1, \ldots, \alpha_I\}$. Then

$$V(x) \geq \sum_i \alpha_i \cdot w_i - \sum_j \max_i \alpha_i \cdot v_j > \max_i \alpha_i \cdot u = \max_{\alpha \in x} u(\alpha).$$

But this contradicts DFC. ▌

We now prove Theorem 5. The necessity of $\succ$ having a finite additive EU representation is obvious. For necessity of DFC, suppose $\succ$ has a weak temptation representation. For any menu x and any i, let α_i denote a maximizer of $a_i u(\beta) + \sum_{j\in J_i} v_j(\beta)$ over $\beta \in x$. Then

$$\begin{aligned} V_w(x) &= \textstyle\sum_i [a_i u(\alpha_i) + \sum_{j\in J_i} v_j(\alpha_i)] - \sum_i \sum_{j\in J_i} \max_{\beta\in x} v_j(\beta) \\ &\leq \textstyle\sum_i [a_i u(\alpha_i) + \sum_{j\in J_i} v_j(\alpha_i)] - \sum_i \sum_{j\in J_i} v_j(\alpha_i) \\ &= \textstyle\sum_i a_i u(\alpha_i) \\ &\leq \max_{\beta\in x} u(\beta). \end{aligned}$$

Hence DFC must hold.

For sufficiency, let V denote a finite additive EU representation of $\succ$. By Lemma 3, we can write this as

$$\begin{aligned} V(x) &= \textstyle\sum_i \max_{\beta\in x} [a_i u(\beta) + \sum_j b_{ij} v_j(\beta) + c_i] - \sum_j \max_{\beta\in x} v_j(\beta) \\ &= \textstyle\sum_i \max_{\beta\in x} [a_i u(\beta) + \sum_j b_{ij} v_j(\beta)] - \sum_j \max_{\beta\in x} v_j(\beta) + \sum_i c_i \end{aligned}$$

where $u(\beta) = V(\{\beta\})$. But

$$u + \sum_j v_j = \sum_i w_i = \sum_i a_i u + \sum_i \sum_j b_{ij} v_j + \sum_i c_i \mathbf{1}.$$

Since $\sum_i a_i = \sum_i b_{ij} = 1$ for all j, this says

$$u + \sum_j v_j = u + \sum_j v_j + \sum_i c_i \mathbf{1},$$

so $\sum_i c_i = 0$.

Let I_+ denote the set of i such that $a_i > 0$. For each $i \in I_+$, let $q_i = a_i$. Let K denote the number of (i,j) pairs for which $b_{ij} > 0$. For each such (i,j), let $k(i,j)$ denote a distinct element of $\{1,\ldots,K\}$. For each $i \in I_+$ and each j such that $b_{ij} > 0$, define a utility function $\hat{v}_{k(i,j)} = [b_{ij}/a_i] v_j$ and let $k(i,j) \in J_i$. For each $i \notin I_+$ and each j with $b_{ij} > 0$, define a utility function $\hat{v}_{k(i,j)} = b_{ij} v_j$ and let $k(i,j) \in J_i$. So for $i \in I_+$,

$$w_i = a_i u + \sum_j b_{ij} v_j = q_i [u + \sum_{j\in J_i} \hat{v}_j].$$

For $i \notin I_+$,

$$w_i = \sum_j b_{ij} v_j = \sum_{j\in J_i} \hat{v}_j.$$

Also,

$$\begin{aligned} \textstyle\sum_j \max_{\beta\in x} v_j(\beta) &= \textstyle\sum_j \sum_i b_{ij} \max_{\beta\in x} v_j(\beta) \\ &= \textstyle\sum_{i\in I_+} \sum_{j\in J_i} q_i \max_{\beta\in x} \hat{v}_j(\beta) + \sum_{i\notin I_+} \sum_{j\in J_i} \max_{\beta\in x} \hat{v}_j(\beta). \end{aligned}$$

Hence

$$V(x) = \sum_{i \in I_+} q_i \max_{\beta \in x} [u(\beta) - c_i(\beta, x)] + \sum_{i \notin I_+} \max_{\beta \in x} [-c_i(\beta, x)]$$

where

$$c_i(\beta, x) = \left[\sum_{j \in J_i} \max_{\beta' \in x} \hat{v}_j(\beta') \right] - \sum_{j \in J_i} \hat{v}_j(\beta).$$

Hence V is a weak temptation representation. ▌

C Proof of Theorem 4

First, we show necessity. Obviously, if $\succ$ has a temptation representation, it has a weak temptation representation, so DFC and existence of a finite additive EU representation are necessary. Hence the following lemma completes the proof of necessity.

Let

$$B(x) = \{\alpha \in x \mid \{\alpha\} \succeq \{\alpha'\},\ \forall \alpha' \in x\}.$$

Lemma 4 *If $\succ$ has a temptation representation, then it satisfies AIT.*

Fix $\succ$ and a temptation representation, V_T. Let β be an approximate improvement for x. Fix any $y \subseteq x$ and $\alpha \in y$ such that $\{\alpha\} \succ \{\beta\}$. Without loss of generality, assume $\alpha \in B(y)$. (If no such x, β, y, and α exist, AIT holds trivially.) By definition of an approximate improvement, there exists a sequence β_n converging to β such that $x \cup \{\beta_n\} \succ x$ for all n.

For any menu z, we can write

$$V_T(z) = \sum_i q_i \max_{\gamma \in z} \left[u(\gamma) + \sum_{j \in J_i} v_j(\gamma) \right] - \sum_i q_i \sum_{j \in J_i} \max_{\gamma \in z} v_j(\gamma).$$

Clearly, then, the fact that $V_T(x \cup \{\beta_n\}) > V_T(x)$ implies that for each n, there is some i with

$$u(\beta_n) + \sum_{j \in J_i} v_j(\beta_n) > \max_{\gamma \in x} \left[u(\gamma) + \sum_{j \in J_i} v_j(\gamma) \right].$$

Otherwise, all the maximized terms in the first sum would be the same at $z = x$ as at $z = x \cup \{\beta_n\}$, while the terms being subtracted off must be at least as large at $z = x \cup \{\beta_n\}$

as at $z = x$. Let i_n^* denote any such i. Because there are finitely many i's, we can choose a subsequence so that i_n^* is independent of n. Hence we can let $i^* = i_n^*$ for all n. Hence

$$u(\beta_n) + \sum_{j \in J_{i^*}} v_j(\beta_n) > \max_{\gamma \in x} \left[u(\gamma) + \sum_{j \in J_{i^*}} v_j(\gamma) \right]$$

for all n, implying

$$u(\beta) + \sum_{j \in J_{i^*}} v_j(\beta) \geq \max_{\gamma \in x} \left[u(\gamma) + \sum_{j \in J_{i^*}} v_j(\gamma) \right].$$

Clearly, then, since $y \subseteq x$,

$$u(\beta) + \sum_{j \in J_{i^*}} v_j(\beta) \geq \max_{\gamma \in y} \left[u(\gamma) + \sum_{j \in J_{i^*}} v_j(\gamma) \right].$$

Subtract $\sum_{j \in J_{i^*}} \max_{\gamma \in y \cup \{\beta\}} v_j(\gamma)$ from both sides to obtain

$$u(\beta) - c_{i^*}(\beta, y \cup \{\beta\}) \geq \max_{\gamma \in y} \left[u(\gamma) - c_{i^*}(\gamma, y \cup \{\beta\}) \right]$$

where c_{i^*} is the self control cost for state i^* from the temptation representation.

Recall that $\alpha \in B(y)$. Hence we have

$$\begin{array}{rcl} V_T(y \cup \{\beta\}) & = & \sum_i q_i \max_{\gamma \in y \cup \{\beta\}} \left[u(\gamma) - c_i(\gamma, y \cup \{\beta\}) \right] \\ & = & q_{i^*} [u(\beta) - c_{i^*}(\beta, y \cup \{\beta\})] + \sum_{i \neq i^*} q_i \max_{\gamma \in y \cup \{\beta\}} \left[u(\gamma) - c_i(\gamma, y \cup \{\beta\}) \right] \\ & \leq & q_{i^*} [u(\beta) - c_{i^*}(\beta, y \cup \{\beta\})] + \sum_{i \neq i^*} q_i \max_{\gamma \in y \cup \{\beta\}} u(\gamma) \\ & = & q_{i^*} [u(\beta) - c_{i^*}(\beta, y \cup \{\beta\})] + (1 - q_{i^*}) u(\alpha) \\ & \leq & q_{i^*} u(\beta) + (1 - q_{i^*}) u(\alpha) \\ & < & u(\alpha) \end{array}$$

where the two weak inequalities follow from $c_i(\gamma, y \cup \{\beta\}) \geq 0$ and the strict inequality follows from $q_{i^*} > 0$ and $\{\alpha\} \succ \{\beta\}$. Hence $\{\alpha\} \succ y \cup \{\beta\}$, so AIT is satisfied. ▌

Turning to sufficiency, for the rest of this proof, let $\succ$ denote a preference with a finite additive EU representation V which satisfies DFC and AIT.

Before moving to the main part of the proof of sufficiency, we get some special cases out of the way. First, it is easy to see that if $\succ$ has a finite additive EU representation, then it has such a representation which is nonredundant in the sense that no w_i or v_j is a constant function and no two states correspond to the same preference over $\Delta(B)$. On the other hand, this nonredundant representation could have $I = 0$, $J = 0$, or both. We first handle these cases, then subsequently focus on the case where $I \geq 1$, $J \geq 1$, no state is a constant preference, and no two states have the same preference over lotteries.

If $I = J = 0$, the preference is trivial in the sense that $x \sim x'$ for all x and x'. In this case, the preference is obviously represented by the temptation representation

$$V(x) = \max_{\beta \in x}[u(\beta) + v(\beta)] - \max_{\beta \in x} v(\beta)$$

where v and u are constant functions. If $I = 0$ but $J \geq 1$, then we have

$$V(x) = K - \sum_j \max_{\beta \in x} v_j(\beta)$$

for an arbitrary constant K. Let w_1 denote a constant function equal to K and define $u = w_1 - \sum_j v_j$. Then

$$V(x) = \max_{\beta \in x}[u(\beta) + \sum_j v_j(\beta)] - \sum_j \max_{\beta \in x} v_j(\beta),$$

giving a temptation representation. Finally, suppose $J = 0$. To satisfy DFC, we must then have $I = 1$, so $V(x) = \max_{\beta \in x} w_1(\beta) + K$ for an arbitrary constant K. Let v_1 be a constant function equal to K and define $u = w_1 - v_1$. Then obviously

$$V(x) = \max_{\beta \in x}[u(\beta) + v_1(\beta)] - \max_{\beta \in x} v_1(\beta),$$

giving a temptation representation.

The remainder of the proof shows the result for the case where the finite additive EU representation has $I \geq 1$ positive states and $J \geq 1$ negative states, none of which are constant and no two of which correspond to the same preference over menus. Following GP, we refer to this as a *regular* representation.

Recall that $B(x)$ is the set of $\alpha \in x$ such that $\{\alpha\} \succeq \{\alpha'\}$ for all $\alpha' \in x$. Define a menu x to be *temptation–free* if there is an $\alpha \in B(x)$ such that $\{\alpha\} \sim x$. That is, x is temptation–free if and only if it is not tempted.

Lemma 5 *Suppose $\succ$ satisfies AIT and has a regular, finite additive EU representation given by*

$$V(x) = \sum_i \max_{\beta \in x} w_i(\beta) - \sum_j \max_{\beta \in x} v_j(\beta).$$

Fix any interior β and any x such that $x \cup \{\beta\}$ is temptation–free and $\beta \notin B(x \cup \{\beta\})$. Then there is no i with

$$w_i(\beta) = \max_{\alpha \in x \cup \{\beta\}} w_i(\alpha).$$

Proof. Suppose not. Suppose there is an interior β, an x such that $x \cup \{\beta\}$ is temptation–free and $\beta \notin B(x \cup \{\beta\})$, and an i with

$$w_i(\beta) = \max_{\alpha \in x \cup \{\beta\}} w_i(\alpha).$$

Because $\beta \notin B(x \cup \{\beta\})$, we know that $u(\beta) < \max_{\alpha \in x} u(\alpha)$, where u is defined by $u(\gamma) = V(\{\gamma\})$ as usual. By hypothesis, the additive EU representation is regular so w_i is not constant. Because w_i is not constant and β is interior, for any $\varepsilon > 0$, we can find a $\hat{\beta}$ within an ε neighborhood of β such that $w_i(\hat{\beta}) > w_i(\beta)$. Hence $w_i(\hat{\beta}) > \max_{\alpha \in x} w_i(\alpha)$. Obviously, if ε is sufficiently small, we will have $u(\hat{\beta})$ close to $u(\beta)$ and hence $u(\hat{\beta}) < \max_{\alpha \in x} u(\alpha)$.

Let $\hat{J}$ denote the set of j such that

$$\max\{v_j(\beta), v_j(\hat{\beta})\} > \max_{\alpha \in x} v_j(\alpha).$$

For each $j \in \hat{J}$, we can find a γ_j such that $v_j(\gamma_j) > v_j(\beta)$ and $w_j(\gamma_j) < w_i(\beta)$. To see that this must be possible, note that the selection of j implies that w_i and $-v_j$ do not represent the same preference. By hypothesis, the additive EU representation is regular so w_i and v_j do not represent the same preference and neither is constant. Hence the v_j indifference curve through β must have a nontrivial intersection with the w_i indifference curve through β. Hence such a γ_j must exist.

Let x' denote the collection of these γ_j's. (If $\hat{J} = \emptyset$, then $x' = \emptyset$.) Let $\beta_\lambda = \lambda\beta + (1-\lambda)\hat{\beta}$. By construction, for all $\lambda \in (0,1)$, w_i ranks β_λ strictly above any $\alpha \in x$. Also, since $w_i(\beta) > w_i(\gamma_j)$ for all j, there is a $\bar{\lambda} \in (0,1)$ such that $w_i(\beta_\lambda) > w_i(\gamma_j)$ for all j for all $\lambda \in (\bar{\lambda}, 1)$. Also, for every $j \notin \hat{J}$, v_j ranks some point in x (and hence in $x' \cup x$) at least weakly above both β and $\hat{\beta}$ and hence above β_λ. Finally, for every $j \in \hat{J}$, $v_j(\gamma_j) > v_j(\beta)$. Hence there is a $\bar{\lambda}' \in (0,1)$ such that $v_j(\gamma_j) > v_j(\beta_\lambda)$ for all $j \in \hat{J}$ and all $\lambda \in (\bar{\lambda}', 1)$. Let $\lambda^* = \max\{\bar{\lambda}, \bar{\lambda}'\}$. For $\lambda \in (\lambda^*, 1)$, then,

$$w_i(\beta_\lambda) > \max_{\alpha \in x' \cup x} w_i(\alpha)$$

$$v_j(\beta_\lambda) < \max_{\alpha \in x' \cup x} v_j(\alpha), \quad \forall j$$

Hence

$$V(x' \cup x \cup \{\beta_\lambda\}) = w_i(\beta_\lambda) + \sum_{k \neq i} \max_{\alpha \in x' \cup x \cup \{\beta_\lambda\}} w_k(\alpha) - \sum_j \max_{\alpha \in x' \cup x} v_j(\alpha).$$

Since the w_i comparison of β_λ to any $\alpha \in x$ or any γ_j is strict, this expression is

$$> \max_{\alpha \in x' \cup x} w_i(\alpha) + \sum_{k \neq i} \max_{\alpha \in x' \cup x \cup \{\beta_\lambda\}} w_k(\alpha) - \sum_j \max_{\alpha \in x' \cup x} v_j(\alpha).$$

Obviously, this is

$$\geq \sum_k \max_{\alpha \in x' \cup x} w_k(\alpha) - \sum_j \max_{\alpha \in x' \cup x} v_j(\alpha) = V(x' \cup x).$$

Hence $x' \cup x \cup \{\beta_\lambda\} \succ x' \cup x$ for all $\lambda \in (\lambda^*, 1)$. Since $\beta_\lambda \to \beta$ as $\lambda \to 1$, this implies β is an approximate improvement for $x' \cup x$. But then AIT implies that $x \cup \{\beta\}$ cannot be temptation–free, a contradiction. ▌

To complete the proof of Theorem 4, we use the following result from Rockafellar [1970] (Theorem 22.2, pages 198–199):

Lemma 6 *Let $z_i \in \mathbf{R}^N$ and $Z_i \in \mathbf{R}$ for $i = 1, \ldots, m$ and let ℓ be an integer, $1 \leq \ell \leq m$. Assume that the system $z_i \cdot y \leq Z_i$, $i = \ell + 1, \ldots, m$ is consistent. Then one and only one of the following alternatives holds:*

(a) There exists a vector y such that

$$z_i \cdot y < Z_i, \ i = 1, \ldots, \ell$$

$$z_i \cdot y \leq Z_i, \ i = \ell + 1, \ldots, m$$

(b) There exist non–negative real numbers $\lambda_1, \ldots, \lambda_m$ such that at least one of the numbers $\lambda_1, \ldots, \lambda_\ell$ is not zero, and

$$\sum_{i=1}^{m} \lambda_i z_i = 0$$

$$\sum_{i=1}^{m} \lambda_i Z_i \leq 0.$$

It is easy to use this result to show that if we have some equality constraints, we simply drop the requirement that the corresponding λ's are non–negative.

Fix $\succ$ with a regular finite additive EU representation which satisfies DFC and AIT. We use Lemma 6 to show that there exists $a_1, \ldots, a_I$, $b_{11}, \ldots, b_{IJ}$, and $c_1, \ldots, c_I$ such that

$$a_i u + \sum_j b_{ij} v_j + c_i \mathbf{1} = w_i, \ \forall i$$

$$\sum_i a_i = 1$$

$$\sum_i b_{ij} = 1, \ \forall j$$

$$-b_{ij} \leq 0, \ \forall i, j$$

$$-a_i < 0, \ \forall i.$$

Because DFC implies that a weak temptation representation exists, the part of the system with only weak inequality constraints is obviously consistent. To state the alternatives implied by the lemma in the most straightforward way possible, let λ_{ik} denote the real number corresponding to the equation

$$a_i u(k) + \sum_j b_{ij} v_j(k) + c_i = w_i(k)$$

where k denotes the kth pure outcome. We use $\bar{\mu}$ to correspond to the equation $\sum_i a_i = 1$, μ_j for the equation $\sum_i b_{ij} = 1$, φ_{ij} for $-b_{ij} \leq 0$, and ψ_i for $-a_i < 0$. Hence Lemma 6 implies that either the a_i's, b_{ij}'s, and c_i's exists or there exists λ_{ik}, $\bar{\mu}$, μ_j, φ_{ij}, and ψ_i such that

$$\varphi_{ij} \geq 0, \quad \forall i, j$$

$$\psi_i \geq 0, \quad \forall i, \text{ strictly for some } i$$

$$\sum_k \lambda_{ik} u(k) + \bar{\mu} - \psi_i = 0, \ i = 1, \ldots, I$$

$$\sum_k \lambda_{ik} v_j(k) + \mu_j - \varphi_{ij} = 0, \ i = 1, \ldots, I; j = 1, \ldots, J$$

$$\sum_k \lambda_{ik} = 0, \ i = 1, \ldots, I$$

$$\sum_i \sum_k \lambda_{ik} w_i(k) + \bar{\mu} + \sum_j \mu_j \leq 0$$

Assume, then, that no a_i's, b_{ij}'s, and c_i's exist satisfying the conditions postulated. Then by Lemma 6, there must be a solution to this system of equations. Note that we cannot have a solution to these equations with $\lambda_{ik} = 0$ for all i and k. To see this, note that the third equation would then imply $\bar{\mu} = \psi_i$ for all i and hence $\bar{\mu} > 0$. Also, from the fourth equation, we would have $\mu_j = \varphi_{ij}$ and hence $\mu_j \geq 0$ for all j. But then the last equation gives $\bar{\mu} + \sum_j \mu_j \leq 0$, a contradiction. Since $\sum_k \lambda_{ik} = 0$, this implies $\max_{i,k} \lambda_{ik} > 0$. Without loss of generality, then, we can assume that $\lambda_{ik} < 1/n$ for all i and k. (Recall that there are n pure outcomes.) Otherwise, we can divide through all equations by $2n \max_{i,k} |\lambda_{ik}|$ and redefine all variables appropriately.

Rearranging the equations gives

$$\sum_k \lambda_{ik} u(k) + \bar{\mu} = \psi_i \geq 0, \ \forall i \text{ with strict inequality for some } i$$

$$\sum_k \lambda_{ik} v_j(k) + \mu_j = \varphi_{ij} \geq 0, \ \forall i, j$$

$$\sum_i \sum_k \lambda_{ik} w_i(k) + \bar{\mu} + \sum_j \mu_j \leq 0$$

For each i, define an interior probability distribution α_i by $\alpha_i(k) = (1/n) - \lambda_{ik}$. Because $\lambda_{ik} < 1/n$ for all i and k, we have $\alpha_i(k) > 0$ for all i and k. Also, $\sum_k \alpha_i(k) = 1 - \sum_k \lambda_{ik} = 1$. Letting β denote the probability distribution $(1/n, \ldots, 1/n)$, we can rewrite the above as

$$u(\beta) + \bar{\mu} \geq u(\alpha_i), \ \forall i \text{ with strict inequality for some } i$$

$$v_j(\beta) + \mu_j \geq v_j(\alpha_i), \quad \forall i$$

$$\sum_i w_i(\beta) + \bar{\mu} + \sum_j \mu_j \leq \sum_i w_i(\alpha_i).$$

The first inequality implies

$$u(\beta) + \bar{\mu} \geq \max_i u(\alpha_i) \tag{1}$$

with a strict inequality for some i. The second inequality implies

$$\sum_j v_j(\beta) + \sum_j \mu_j \geq \sum_j \max_i v_j(\alpha_i). \tag{2}$$

Turning to the third inequality, recall that $\sum_i w_i = u + \sum_j v_j$. Hence the third inequality is equivalent to

$$u(\beta) + \sum_j v_j(\beta) + \bar{\mu} + \sum_j \mu_j \leq \sum_i w_i(\alpha_i).$$

Summing equations (1) and (2) yields

$$u(\beta) + \sum_j v_j(\beta) + \bar{\mu} + \sum_j \mu_j \geq \max_i u(\alpha_i) + \sum_j \max_i v_j(\alpha_i)$$

so

$$\sum_i w_i(\alpha_i) - \sum_j \max_i v_j(\alpha_i) \geq u(\beta) + \sum_j v_j(\beta) + \bar{\mu} + \sum_j \mu_j - \sum_j \max_i v_j(\alpha_i) \geq \max_i u(\alpha_i). \tag{3}$$

Let $x = \{\alpha_1, \ldots, \alpha_I\}$. Then

$$V(x) \geq \sum_i w_i(\alpha_i) - \sum_j \max_i v_j(\alpha_i) \geq \max_i u(\alpha_i).$$

By DFC, $\max_i u(\alpha_i) \geq V(x)$. Hence

$$V(x) = \sum_i w_i(\alpha_i) - \sum_j \max_i v_j(\alpha_i) = \max_i u(\alpha_i).$$

Hence x is a temptation–free menu. Note that the first equality in the last equation implies that α_i maximizes w_i for all i. Also, the second equality together with equation (3) implies that the weak inequalities in equations (1) and (2) must be equalities. In particular, then,

$$u(\beta) + \bar{\mu} = \max_i u(\alpha_i).$$

However, recall that

$$u(\beta) + \bar{\mu} \geq u(\alpha_i), \ \forall i \text{ with strict inequality for some } i$$

That is, there must be some k for which $u(\alpha_k) < \max_i u(\alpha_i)$. Hence $x \neq B(x)$. But α_i maximizes w_i for every i, contradicting Lemma 5.

Hence there must exist such a_i, b_{ij}, and c_i. It is easy to use the proof of Theorem 5 to complete the construction of a temptation representation. ▌

D Proof of Lemma 1

Proof. (Necessity.) We show that if $\succ$ has a finite additive EU representation with only one positive state and $x \succeq y$, then $x \succeq x \cup y$. It is not hard to see that

$$V(x \cup y) = \sum_i \max\left\{\max_{\beta \in x} w_i(\beta), \max_{\beta \in y} w_i(\beta)\right\} - \sum_j \max\left\{\max_{\beta \in x} v_j(\beta), \max_{\beta \in y} v_j(\beta)\right\}.$$

When there is only one positive state, $I = 1$, so we can rewrite this as

$$\begin{array}{rcl} V(x \cup y) & = & \max\left\{\max_{\beta \in x} w_1(\beta), \max_{\beta \in y} w_1(\beta)\right\} \\ & & -\sum_j \max\left\{\max_{\beta \in x} v_j(\beta), \max_{\beta \in y} v_j(\beta)\right\}. \end{array}$$

Hence

$$\begin{array}{rcl} V(x \cup y) & \leq & \max\left\{\max_{\beta \in x} w_1(\beta), \max_{\beta \in y} w_1(\beta)\right\} \\ & & \quad - \max\left\{\sum_j \max_{\beta \in x} v_j(\beta), \sum_j \max_{\beta \in y} v_j(\beta)\right\} \\ & \leq & \max\left\{\max_{\beta \in x} w_1(\beta) - \sum_j \max_{\beta \in x} v_j(\beta), \right. \\ & & \qquad \left. \max_{\beta \in y} w_1(\beta) - \sum_j \max_{\beta \in y} v_j(\beta)\right\} \\ & = & \max\{V(x), V(y)\} = V(x). \end{array}$$

Hence $x \succeq x \cup y$.

(Sufficiency.) Suppose $\succ$ has a finite additive EU representation and satisfies positive set betweenness. Assume, contrary to our claim, that this representation has more than one positive state. So $\succ$ has a representation of the form

$$V(x) = \sum_{i=1}^{I} \max_{\beta \in x} w_i(\beta) - \sum_{j=1}^{J} \max_{\beta \in x} v_j(\beta)$$

where $I \geq 2$. Without loss of generality, we can assume that w_1 and w_2 represent different preferences over $\Delta(B)$ — otherwise, we can rewrite the representation to combine these two states into one. Let $\hat{x}$ denote a sphere in the interior of $\Delta(B)$. Let

$$x = \left[\bigcap_{i=1}^{I} \{\beta \in \Delta(B) \mid w_i(\beta) \leq \max_{\beta' \in \hat{x}} w_i(\beta')\}\right] \bigcap \left[\bigcap_{j=1}^{J} \{\beta \in \Delta(B) \mid v_j(\beta) \leq \max_{\beta' \in \hat{x}} v_j(\beta')\}\right].$$

Because $\hat{x}$ is a sphere and because I and J are finite, there must be a w_i indifference curve which makes up part of the boundary of x for $i = 1, 2$. Fix a small $\varepsilon > 0$. For $i = 1, 2$ and $k = 1, \ldots, I$, let $\varepsilon_k^i = 0$ for $k \neq i$ and $\varepsilon_i^i = \varepsilon$. Finally, for $i = 1, 2$, let y_i equal

$$\left[\bigcap_{k=1}^{I} \{\beta \in \Delta(B) \mid w_k(\beta) \leq \max_{\beta' \in \hat{x}} w_k(\beta') - \varepsilon_k^i\}\right] \bigcap \left[\bigcap_{j=1}^{J} \{\beta \in \Delta(B) \mid v_j(\beta) \leq \max_{\beta' \in \hat{x}} v_j(\beta')\}\right].$$

Because I and J are finite, if ε is sufficiently small,

$$\max_{\beta\in y_i} w_k(\beta) = \max_{\beta\in x} w_k(\beta), \quad \forall k \neq i$$

and

$$\max_{\beta\in y_i} v_j(\beta) = \max_{\beta\in x} v_j(\beta), \quad \forall j.$$

Hence $x \sim y_1 \cup y_2$. Also,

$$\max_{\beta\in y_i} w_i(\beta) < \max_{\beta\in x} w_i(\beta).$$

Hence $x \succ y_i$, $i = 1, 2$. Hence $y_1 \cup y_2 \succ y_i$, $i = 1, 2$, contradicting positive set betweenness. ▌

E Proof of Theorem 7

Proof. Necessity is obvious. For sufficiency, assume $\succ$ has a finite additive EU representation and satisfies DFC and negative set betweenness. We know from Lemma 2 that it has only one negative state. Using this and Lemma 3, we see that $\succ$ can be represented by a function V of the form

$$V(x) = \sum_{i=1}^{I} \max_{\beta\in x}[a_i u(\beta) + b_i v(\beta)] - \max_{\beta\in x} v(\beta)$$

where $a_i \geq 0$ and $b_i \geq 0$ for all i and $\sum_i a_i = \sum_i b_i = 1$. (The argument in the proof of Theorem 5 showing that $\sum_i c_i = 0$ applies here as well.)

We can assume without loss of generality that $a_i > 0$ for all i. To see this, suppose $a_1 = 0$. Then we can write

$$V(x) = \sum_{i=2}^{I} \max_{\beta\in x}[a_i u(\beta) + b_i v(\beta)] - \max_{\beta\in x}(1 - b_1)v(\beta).$$

If $b_1 = 1$, then $b_i = 0$ for all $i \neq 1$. Because $a_1 = 0$ and $\sum_i a_i = 1$, we then have $V(x) = \max_{\beta\in x} u(\beta)$. This is a V_{US} representation with $I = 1$ and $\gamma_1 = 0$. So suppose $b_1 < 1$. Let $\hat{v} = (1 - b_1)v$ and for $i = 2, \ldots, I$, let $\hat{b}_i = b_i/(1 - b_1)$. Note that $\sum_{i=2}^{I} \hat{b}_i = 1$. Hence we can rewrite V as

$$V(x) = \sum_{i=2}^{I} \max_{\beta\in x}[a_i u(\beta) + \hat{b}_i \hat{v}(\beta)] - \max_{\beta\in x} \hat{v}(\beta).$$

Continuing as needed, we can eliminate any i with $a_i = 0$.

Given that $a_i > 0$ for all i, let $q_i = a_i$ and let $\gamma_i = b_i/a_i$. With this change of notation, V can be rewritten in the form of V_{US}. ▌

References

[1] Dekel, E., B. Lipman, and A. Rustichini, "Representing Preferences with a Unique Subjective State Space," *Econometrica*, **69**, July 2001, 891–934.

[2] Gul, F., and W. Pesendorfer, "Temptation and Self–Control," *Econometrica*, **69**, November 2001, 1403–1435.

[3] Harsanyi, J., "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," *Journal of Political Economy*, **63**, 1955, 309–321.

[4] Kreps, D., "A Representation Theorem for 'Preference for Flexibility'," *Econometrica,* **47**, May 1979, 565–576.

[5] Kreps, D., "Static Choice and Unforeseen Contingencies" in *Economic Analysis of Markets and Games: Essays in Honor of Frank Hahn*, P. Dasgupta, D. Gale, O. Hart, and E. Maskin, eds., Cambridge, MA: MIT Press, 1992, 259–281.

[6] Machina, M., "Dynamic Consistency and Non–Expected Utility Models of Choice Under Uncertainty," *Journal of Economic Literature*, **27**, 1989, 1622–1668.

[7] Noor, J., "Temptation, Welfare, and Revealed Preference," University of Rochester working paper, 2004.

[8] Rockafellar, R. T., *Convex Analysis*, Princeton, NJ: Princeton University Press, 1970.

[9] Weymark, J., "A Reconsideration of the Harsanyi–Sen Debate on Utilitarianism," in J. Elster and J. Roemer, eds., *Interpersonal Comparisons of Well–Being*, Cambridge: Cambridge University Press, 1991, 255–320.