Open access • Journal Article • DOI:10.1111/J.1467-937X.2009.00560.X

Temptation-Driven Preferences — Source link

Eddie Dekel, Barton L. Lipman, Aldo Rustichini

Institutions: Tel Aviv University, Boston University, University of Cambridge

Published on: 01 Jul 2009 - The Review of Economic Studies (Oxford University Press)

Topics: Temptation, Rational agent and Axiom

Temptation–Driven Preferences[1]

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

April 2006
Current Draft

[1]We thank Drew Fudenberg, Fabio Maccheroni, Massimo Marinacci, Jawwad Noor, Ben Polak, Phil Reny and numerous seminar audiences for helpful comments. We also thank the NSF for financial support for this research. We particularly thank Todd Sarver for comments and for agreeing to the move of Theorem 6 from our joint paper (Dekel, Lipman, Rustichini, and Sarver [2005]) to this one.

[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 including some which may be important in applications. We propose a pair of alternative axioms called DFC, *desire for commitment*, and AIC, *approximate improvements are chosen*. 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. AIC is based on the idea that if adding an option to a menu improves the menu, it is because that option is chosen under some circumstances. From this interpretation, the axiom concludes that if an improvement is worse (as a commitment) than some commitment from the menu, then the best commitment from the menu is strictly preferred to facing the menu. We show that these axioms characterize a natural generalization of the Gul–Pesendorfer representation.

1 Introduction

What potentially observable 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 current view of what actions she would like to choose, but knows that at the time these choices are to be made, she will be pulled by conflicting desires. For clarity, we refer to her current view of desirable actions as her *commitment preference* since this describes the actions she would commit herself to if possible. We interpret and frequently discuss this preference as the agent's view of what is normatively appropriate, though this is not a formal part of the model.[1] We refer to the future desires that may conflict with the commitment preference as *temptations*. We view this conflict as independent of the set of feasible options in the sense that whether one item is more tempting than another is independent of what other options are available. Thus we do impose a certain structure on the way temptation affects the agent. Also, we allow the possibility that the extent or nature of temptation is random, but do not allow similar randomness regarding what is normatively preferred. While there is undoubtedly an element of arbitrariness in this modeling choice, we choose to rule out uncertainty about what is normatively preferred to separate temptation–driven behavior from the desire for flexibility which such uncertainty would generate. We retain uncertainty about temptation for two reasons.[2] First, as we will see, some behavior which is very intuitive as an outcome of temptation is (unexpectedly) difficult to explain without uncertainty about temptation. Second, we believe uncertainty about temptations is likely to be important in applications.[3]

Our approach builds on earlier work by Gul–Pesendorfer [2001] (henceforth GP) and Dekel–Lipman–Rustichini [2001] (DLR). DLR consider a rather general model of preferences over menus, from which choice is made at a later date. (A menu can be interpreted either literally or as an action which affects subsequent opportunities.) DLR show that preferences over menus can be used to identify an agent's subjective beliefs regarding her future tastes and behavior. The set of preferences considered by DLR can be interpreted as allowing for a desire for flexibility, concerns over temptation, or both considerations, as well as preferences with entirely different interpretations.[4]

GP were the first to use preferences over menus to study temptation. To see the

[1]See Noor [2006a] for a critique of such interpretations.

[2]Also, allowing uncertainty about normative preferences poses severe identification problems. See Section 6 for details.

[3]It is true that uncertainty about what is normatively appropriate may also be important in applications as well; see Amador, Werning, and Angeletos [2006].

[4]For examples of different motivations, see Sarver [2005] or Ergin and Sarver [2005].

intuition for how this works, recall that temptation refers to desires to deviate from the commitment preference. The commitment preference is naturally identified as the preference over singleton menus, since such menus correspond exactly to commitments to particular choices. Thus temptation can be identified by seeing how preferences over non–singleton menus differ from what would be implied by the commitment preferences if there were no temptation. That is, if $\{a\} \succ \{b\}$, so the agent prefers a commitment of a to a commitment of b, then if there were no temptation (or other "nonstandard" motives), we would have $\{a, b\} \sim \{a\}$ since she would choose a from $\{a, b\}$. With temptation, though, $\{a\}$ may be strictly preferred to $\{a, b\}$.

Using this intuition, GP focus on temptation alone by adding a *set betweenness* axiom to the DLR model. As we explain in more detail in subsequent sections, this axiom has the implication that temptation is one–dimensional in the sense that for any menu, temptation only affects the agent through the "most tempting" item on the menu. This rules out many intuitive kinds of temptation–driven behavior. For example, it rules out uncertainty about temptation where the agent cannot be sure which item on a menu will be the most tempting one or how strong the temptation will be. We give examples in Section 3 to illustrate such possibilities.

We believe that uncertainty about temptation is important for applications. In reality, an agent cannot easily "fine tune" her commitments. That is, it is difficult to find a way to commit oneself to some exact course of action without allowing any alternative possibilities. Instead, real commitments tend to be costly actions which alter one's incentives to engage in "desired" or "undesired" future behaviors. Casual observation suggests that such commitments often involve overcommitment (spending more *ex ante* to commit to a certain behavior than turns out *ex post* to be necessary) or undercommitment (finding out *ex post* that the change in one's incentives was not sufficient to achieve the desired effect). Neither phenomenon seems consistent with a model without uncertainty.

Similarly, much of the real complexity of achieving commitment comes from the multidimensional character of temptation. To see the point, first suppose the only possible temptation is overspending on current consumption. In this case, the agent can avoid temptation by committing herself to a minimum level of savings. Now suppose there are other temptations that may strike as well, such as the temptation to be lazy and avoid dealing with needed home repairs or other time–consuming expenditures. In this case, the commitment to saving may worsen the agent's ability to deal with other temptations.

As GP argue, it was natural for them to begin the study of temptation by narrowing to a particularly simple version of the phenomenon. Our goal is to use the DLR framework to build on their analysis and carry out the logical next step in the study of temptation, namely identifying the broadest possible set of temptation–driven behavior.

We explore how to model temptation along so as to be able to study behavior that follows from this feature. Just as it is insightful to distinguish between risk and ambiguity, it is instructive to identify behavior due to temptation. Naturally, other factors may lead to similar behavior and we cannot say that the behavior we identify *proves* the agent was tempted, only that it is *consistent* with temptation. But our results do specify what behavior is *not* a consequence of temptation (as we define it). Some of our results also characterize specific forms of temptation, such as uncertainty about the strength of temptation. In the opposite direction, one could extend the form of temptation to allow temptation to depend not only on the tempting item but on the whole set of available items, which would allow for even richer behavior than what we identify.

After simplifying by means of a finiteness axiom, we carry out this next step by adding a pair of axioms to DLR. We show that these axioms characterize a natural generalization of the GP representation. The first axiom, DFC or *desire for commitment*, 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."

The second axiom, AIC or *approximate improvements are chosen*, identifies a circumstance in which the preference for commitment is strict. The key to the axiom is what additional implications we deduce from the fact that adding an option, say β, to a menu x improves the menu. We interpret such an improvement as saying that β is chosen under *some* conditions. Hence, we draw the conclusion that if β is worse than the normatively best $\alpha \in x$, then the agent strictly prefers commitment to α over facing the menu. (The axiom asks for this to holds for perturbations well.) Given the interpretation of the axioms and the intuitive nature of the representation they generate, we conclude that DFC and AIC 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. We explain how these special cases have natural interpretations as restrictions on the kinds of temptation faced by the agent.

As briefly mentioned above, our analysis is based on a simplified version of DLR, the development of which is another contribution of the present paper. To maintain a unified focus, the text focuses almost entirely on the issue of temptation and the Appendix contains a complete explanation of how we add a finiteness requirement to DLR.

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 in which we are interested. In Section 4, we give representation results. Because DFC is a simpler axiom than AIC and because it is a convenient step in the analysis, we also state the representation generated by adding only this axiom. 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, 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.

The basic representation on which we build is what we will call a *finite additive EU representation*. This adds a finite state requirement to what DLR called an additive EU representation. Formally, we say that a utility function over lotteries, $U : \Delta(B) \to \mathbf{R}$ is an expected–utility function if

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

for all β (where $U(b)$ is the utility of the degenerate lottery with probability 1 on b).

Definition 1 *A* finite additive EU representation *is a pair of finite collections of expected–utility functions over $\Delta(B)$, $w_1, \ldots, w_I$ and $v_1, \ldots, v_J$ such that the function*

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

represents $\succ$.

DLR, as modified in the corrigendum (Dekel, Lipman, Rustichini, and Sarver [2005], henceforth DLRS), characterize this class of representations without the finiteness requirement. Theorem 6 in the appendix extends DLR and DLRS by characterizing the set of preferences with a finite additive EU representation.[5]

DLR interpret the different utility functions over $\Delta(B)$ as different states of the world, referring to the I states corresponding to the w_i's as *positive states* and J states

[5]In addition to finiteness, the finite additive EU representation differ from DLR's additive EU representation in three respects. First, DLR included a nonemptiness requirement as part of the definition of an additive EU representation. Consequently, their axiom differ from those of Theorem 6 by including a nontriviality axiom. Second, DLR required that none of the utility functions be redundant. Third, in the infinite case, it is not without loss of generality to have equal weights on all the w_i's and v_j's, so the representation in DLR also specifies a measure on the (infinite) index sets I and J.

corresponding to the v_j's as *negative states*. To understand this interpretation most simply, suppose there are no negative states — i.e., $J = 0$. Then it seems natural to interpret 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 these 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 is the expected value of the maximum. If the w_i's are equally likely, we obtain the value above.[6] This interpretation was originally offered by Kreps [1979, 1992] who first considered preferences over sets as a model of 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 Kreps' monotonicity axiom does this.

Axiom 1 (Monotonicity) *If $x \subset x'$, then $x' \succeq x$.*

It is straightforward to combine results in DLR with Theorem 6 to show the following.[7]

Observation 1 *Assume the preference $\succ$ has a finite additive EU representation. Then $\succ$ has a representation with $J = 0$ if and only if it satisfies monotonicity.*

Intuitively, monotonicity says that the agent always 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 interpret as describing a forward–looking agent with beliefs about her possible future needs.

GP's approach provides an alternative interpretation of the finite additive EU representation by imposing a different restriction on that class of preferences. They recognized that temptation and self–control could be studied using this sets of lotteries framework if one does not impose monotonicity. If the agent anticipates being 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 Observation 1 by replacing monotonicity with an axiom they call *set betweenness*.

[6] The interpretation of the w_i's as equally likely is only for intuition. As is standard with state dependent utility, we can change the probabilities in essentially arbitrary ways and rescale the w_i's to leave the overall utility unchanged. Hence the probabilities cannot be identified. See, however, the third comment in the preceding footnote.

[7] If $\succ$ has a representation with $J = 0$, it will also have other representations with $J > 0$. To see this, note that we can add a v_j satisfying $v_j(\beta) = k$ for all β to any representation and not change the preference being represented. This is why DLR imposed a requirement that no "redundant" states are included. For the purposes of this paper, it is simpler to allow redundancy.

Axiom 2 (Set Betweenness) *If $x \succeq y$, then $x \succeq x \cup y \succeq y$.*

To understand this axiom, consider a dieting agent's choice of a restaurant for lunch where x, y, and $x \cup y$ are the menus 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. Since the agent is dieting, presumably $x \succ y$. Given this, how should the agent rank the menu $x \cup y$ relative to the other two? A natural hypothesis is that the third restaurant would lie 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, $x \cup y$ 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 $x \succeq x \cup y \succeq y$.

The relevant representation in GP is the following.

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 $\succ$.

It is easy to see that this is a finite additive EU representation with one positive state and one negative state where $w_1 = u + v$ and $v_1 = v$. Thus it comes as no surprise that the axioms GP use for this representation include those we use in Theorem 6 to characterize finite additive EU representations.[8] Hence we can paraphrase their result as

Theorem 1 (GP, Theorem 1) *$\succ$ has a self–control representation if and only if it has a finite additive EU representation and satisfies set betweenness.*

To interpret GP's representation, note that u represents the commitment preference — the preference over singletons — as $V_{GP}(\{\beta\}) = u(\beta)$ for any β. For any menu x and any $\beta \in x$, let

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

[8] Specifically, their axioms are the same as those we use in Theorem 6 except that they have set betweenness instead of our finiteness axiom. One can show that set betweenness implies finiteness. It is worth noting that they consider a more general setting than us in that they assume B is compact, not finite.

Intuitively, c is the foregone utility according to v from choosing β from x instead of choosing optimally 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.

To interpret these results, consider the set of preferences with a finite additive EU representation. Intuitively, the subset of these preferences which are monotonic corresponds to those agents who value flexibility but are not affected by temptation. It seems natural to call such preferences *flexibility–driven*, as both the axiom and the representation it generates seem to describe such an agent. In other words, defining flexibility–driven preferences as those which can be explained by flexibility considerations alone, it seems natural to conclude that monotonicity characterizes these preferences.

Analogously, we refer to preferences which can be explained solely by a concern about temptation as *temptation–driven*. It seems natural to say that the preferences that satisfy set betweenness are temptation–driven preferences. However, set betweenness does not appear to be as complete a statement of "temptation–driven preferences" as monotonicity is for "flexibility–driven." In fact, it is not hard to give examples of behavior which seems temptation–driven but which violates set betweenness. This suggests that set betweenness is stronger than 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 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 other representations that may be of interest.

Example 1.

Consider a dieting agent who 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, if the agent 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 craves 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.[9]

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:

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

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)

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

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:

$$V_2(x) = \max_{\beta \in x}[u(\beta) + v_1(\beta) + v_2(\beta)] - \max_{\beta \in x} v_1(\beta) - \max_{\beta \in x} v_2(\beta). \tag{1}$$

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

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

so that $V_2(x) = \max_{\beta \in x}[u(\beta) - c(\beta, x)]$. 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.

We note that there is one odd feature of this representation. If the agent succumbs to either temptation, he still suffers a cost associated with the temptation he does not consume. That is, the self–control cost associated with choosing *either* snack from the menu $\{b,c,p\}$ is 6, not zero. Arguably, it should be feasible for the agent to succumb to temptation and incur no self–control cost. We return to this issue in the conclusion.

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 prefers a chance of sticking to her diet to committing herself to violating it so $\{b,y\} \succ \{y\}$. Also, if the agent cannot avoid having ice cream available, 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. To see why this cannot occur in their model, note that

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

while $V_{GP}(\{y\}) = u(y) = u(y) + v(y) - v(y)$. Obviously, $\max\{v(b), v(y)\} \geq v(y)$. So $V_{GP}(\{b,y\}) > V_{GP}(\{y\})$ 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 $V_{GP}(\{b, i, y\}) \leq V_{GP}(\{b, i\})$. That is, we must have $\{b, i\} \succeq \{b, i, y\}$.[10]

To see this more intuitively, note that $\{b, y\} \succ \{y\}$ says that adding b improves the menu $\{y\}$. As discussed in the introduction, we interpret this as saying that the agent considers it possible that she would choose b from the menu $\{b, y\}$, an interpretation we share with GP. However, in GP, the agent has no uncertainty about temptation, so this statement means she *knows* she will definitely choose b from $\{b, y\}$. Consequently, she will definitely *not* choose y whenever b is available.[11] Hence the only possible effect of adding y to a menu which contains b is to increase self–control costs. Hence GP require $\{b, i, y\} \preceq \{b, i\}$.

This intuition suggests that uncertainty about temptation is critical to rationalizing this preference. The following simple generalization of GP to incorporate uncertainty allows the intuitive preference suggested above. Let

	u	v
b	6	0
i	0	8
y	4	6

and let

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

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. While we do not wish to argue that the axioms needed for such a representation are innocuous, it is not obvious that temptation should require some violation of them (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.

[10]This conclusion does not follow from set betweenness alone. It is easy to give examples of preferences which satisfy set betweenness and avoid this problem but which do not have an additive EU representation.

[11]Note that this conclusion relies on the assumption that temptation does not lead the agent to violate independence of irrelevant alternatives. That is, we are assuming that if the agent would choose b over y from one set, she would never choose y when b is available. See Section 6 for further discussion.

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 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.

We also study one less interpretable representation which is useful as an intermediate step.

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.[12]

4 Results

The following axiom seems to be a natural part of a definition of temptation–driven.

Axiom 3 (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 worse for her diet 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, that is, when there is some $\alpha \in x$ such that $\{\alpha\} \succ x$.

To get some intuition for the axiom we will propose, consider the following example, similar to Example 2, where the three goods are broccoli (b), low fat frozen yogurt (y), and high fat ice cream (i). Assume that $\{b\} \succ \{y\} \succ \{i\}$, so that broccoli is best for the

[12]One way to interpret the weak temptation representation is that it is a limiting case of temptation representations. To see this, fix a weak temptation representation with $I > I'$ and any $\varepsilon \in (0, 1)$. We can define a (strict) temptation representation with I "states" by shifting ε of the probability on the first I' states to the remaining $I - I'$ states, adjusting the cost functions at the same time. More specifically, define $\hat{q}_i = q_i - \varepsilon/(I - I')$ for $i \leq I'$ and $\hat{q}_i = \varepsilon/(I - I')$ for $i = I' + 1, \ldots, I$. For $\varepsilon > 0$ sufficiently small, $\hat{q}_i > 0$ for all i. For $i \leq I'$, let $\hat{c}_i = c_i$. For $i = I' + 1, \ldots, I$, define new cost functions $\hat{c}_i = (1/\hat{q}_i)c_i$. Consider the payoff to any menu as computed by this temptation representation minus the payoff as computed by the original weak representation. It is easy to see that this difference is proportional to ε and so converges to 0 as $\varepsilon \downarrow 0$. In this sense, we have constructed a sequence of temptation representations converging to the weak representation.

agent's diet and ice cream is worst. As we argued earlier, it seems plausible that adding y to the menu $\{b, i\}$ improves the menu since y is a useful compromise when tempted. So assume that $\{b, i, y\} \succ \{b, i\}$. As suggested in the introduction and as we argue at greater length below, if adding an item to a menu improves the menu, this should be interpreted as implying that the added item is sometimes chosen from the menu. That is, we should conclude from $\{b, i, y\} \succ \{b, i\}$ that y is sometimes chosen from the menu $\{b, i, y\}$. Hence with this menu, the agent will sometimes break her diet, choosing y instead of b. Consequently, she should *strictly* prefer committing herself to the broccoli. That is, we should conclude $\{b\} \succ \{b, i, y\}$. In addition, if y is sometimes chosen over b and i, it should also be sometimes chosen from the menu $\{b, y\}$. Thus the dieter sometimes breaks her diet with this menu too, implying $\{b\} \succ \{b, y\}$. These implications are the content of our next axiom when applied to this example: since adding y improves the menu $\{b, i\}$, we require that $\{b\}$ is strictly preferred to both $\{b, i, y\}$ and $\{b, y\}$.

More generally, suppose we have a menu x with the property that adding β to x strictly improves the menu for the agent in the sense that $x \cup \{\beta\} \succ x$. (So think of $x = \{b, i\}$ and $\beta = y$.) In such a case, we say β is *an improvement for* x. How should we interpret this property? In principle, there are many reasons why adding an element to a menu might improve the menu. For example, a menu may be "prettier" with certain lines added to it. An agent may simply like having options, even knowing she would never choose them. More related to temptation, adding a particularly disgusting dessert option might make it easier for a dieter to avoid dessert since reading the menu makes dessert unappetizing.

Our goal is to characterize agents who face temptation but are otherwise "standard rational agents." As such, we consider an agent for whom the items on a menu have a certain appeal which is menu–independent, an appeal which may create internal conflicts which the agent has to resolve. Thus we assume that the normative appeal and the extent of temptation of any given item is independent of the other items in the menu.

In light of this, it seems natural to assume that adding an element to a menu does not make it easier to choose other elements or create value separately from choice (as in the case of a "prettier" menu). That is, the only effect adding an unchosen element can have is to increase self–control costs. With this principle, we interpret $x \cup \{\beta\} \succ x$ as saying that the agent at least considers it possible that she would choose β from the menu $x \cup \{\beta\}$. We emphasize that this is only an interpretation, not a theorem. We are arguing that our definition of temptation strongly suggests this interpretation, not that it "proves" it.[13,14]

[13]Gul and Pesendorfer [2005] also argue for this interpretation of β improving x.

[14]It is difficult if not impossible to draw definitive conclusions about choices the agent would make from a menu based on preferences over menus. At best, we can interpret preferences over menus as suggesting that the agent perceives her future choice in a particular way. For example, consider an agent whose preference over menus has a temptation representation. We interpret the representation as

Under this interpretation of the preference, what else should be true? Suppose α is the best item for her diet in x (i.e., is optimal according to the commitment preference) and $\{\alpha\} \succ \{\beta\}$. (In terms of the example, think of $\alpha = b$.) So $\alpha \in x$ is strictly better for the agent's diet than β and yet she considers it possible she would choose β from the menu $x \cup \{\beta\}$. Thus the agent must consider it possible that her choice from the menu $x \cup \{\beta\}$ is inconsistent with her commitment preference. Hence she should strictly prefer committing herself to α rather than facing the menu $x \cup \{\beta\}$. That is, commitment must be strictly valuable in the sense that $\{\alpha\} \succ x \cup \{\beta\}$.

Similarly, consider some $x' \subseteq x$. (Think of $x' = \{b\}$.) If the agent considers it possible that she would choose β from $x \cup \{\beta\}$, surely she also considers it possible she would choose β from $x' \cup \{\beta\}$.[15] Again, if the best $\alpha \in x'$ for her diet satisfies $\{\alpha\} \succ \{\beta\}$, we should conclude that the agent would strictly prefer the commitment $\{\alpha\}$ to facing the menu $x' \cup \{\beta\}$.

To summarize, we interpret $x \cup \{\beta\} \succ x$ to mean that β is sometimes chosen from $x \cup \{\beta\}$ and hence from $x' \cup \{\beta\}$ for any $x' \subseteq x$. If the best $\alpha \in x'$ satisfies $\{\alpha\} \succ \{\beta\}$, this implies that the agent does not always choose from $x' \cup \{\beta\}$ according to her commitment preferences. Therefore, commitment is strictly valuable for $x' \cup \{\beta\}$ in the sense that $\{\alpha\} \succ x' \cup \{\beta\}$. Since the key to this intuition is that $x \cup \{\beta\} \succ x$ implies β is sometimes chosen from $x \cup \{\beta\}$, we summarize this by saying *improvements are (sometimes) chosen.*

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 AIC, *approximate improvements are chosen.* Formally,

saying that the agent assigns probability q_i to being tempted according to cost function c_i. It seems natural, then, to say that if the agent has menu x, then with probability q_i, the agent will choose a $\beta \in x$ which maximizes $u(\beta) - c_i(\beta, x)$. However, this conclusion is purely an interpretation of the model, not a theorem which can be proven. The only primitive in the model is a preference over menus, so we have no information about choice from the menu with which to confirm this interpretation. GP resolve this problem by extending the preference over menus to menu–choice pairs, but this approach has a severe problem. To state it most simply, let $x = \{a, b, c\}$ and let $\succ^*$ denote this extended preference. Suppose $(x, a) \succ^* (x, b) \succ^* (x, c)$. GP interpret this to say that a is chosen from menu x. While this conclusion seems natural, how are we to interpret $(x, b) \succ^* (x, c)$? There is no choice which can reveal this preference to us. If x is the set of choices available, neither b nor c would be chosen by the agent. Asking the agent to compare (x, b) to (x, c) is like asking the agent which she prefers: being offered x but forced to choose b or being offered x but forced to choose c. In what sense is x the available set if the agent must choose something other than a from the set? Because of these problems, we avoid hypotheses on choice from menus to maximum extent possible.

[15] 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 we discuss in Section 6, this is not necessarily an appropriate assumption for modeling temptation.

define β to be an *approximate improvement for x* if

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

where cl denotes closure. Let $B(x)$ denote the set of best commitments in x. That is,

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

Then we have

Axiom 4 (AIC: Approximate Improvements are Chosen) *If β is an approximate improvement for x, $x' \subseteq x$, and $\alpha \in B(x')$ satisfies $\{\alpha\} \succ \{\beta\}$, then $\{\alpha\} \succ x' \cup \{\beta\}$.*

Theorem 2 *$\succ$ has a temptation representation if and only if it has a finite additive EU representation and satisfies DFC and AIC.*

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 2, we also show

Theorem 3 *$\succ$ 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, we characterize the "no uncertainty" representation V_2 in (1) of Example 1 and the "uncertain strength of temptation" representation V_3 in (2) of Example 2. 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. 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 5 (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 6 (Negative Set Betweenness) $\succ$ *satisfies negative set betweenness if whenever* $x \succeq y$*, we have* $x \cup y \succeq y$*.*

The following lemma characterizes the implication of positive set betweenness.[16]

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.*

[16] See also Kopylov [2005] who gives a generalization to I positive states and J negative states.

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.

Using the change of variables discussed above, this lemma obviously yields

Theorem 4 *$\succ$ 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 1 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

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

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

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

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.

We have

Theorem 5 *$\succ$ 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.

One case of particular interest addresses a potential concern mentioned in the discussion of Example 1. Our general representation allows cost functions that depend on more than one temptation in the sense that we have

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

where J_i need not be a singleton. In general, such a representation will have the property that there is no choice the agent can make which will reduce c_i to 0. One might prefer to assume that if the agent gives in to temptation, the self–control cost is zero. But if J_i is not a singleton and the v_j's are maximized at different points in x, this is impossible, arguably implying that these representations include considerations other than temptation such as regret.[17] This motivates considering a restriction to what we call a *simple representation*, a temptation representation with the property that J_i is a singleton for all i. We conjecture that $\succ$ has a simple representation if and only if it has a finite additive EU representation and satisfies *weak set betweenness*:

Axiom 7 (Weak Set Betweenness) *If $\{\alpha\} \succeq \{\beta\}$ for all $\alpha \in x$ and $\beta \in y$, then $x \succeq x \cup y \succeq y$.*

Another issue which may be of interest is introducing uncertainty about what is normatively desirable as well as about temptation. To some extent, however, this problem is *too* easily solved. More specifically, any finite additive EU representation can be written as a temptation representation with uncertainty about normative preferences. To see the point, return to the general finite additive EU representation where

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

Partition the set $\{1, \ldots, J\}$ into I sets, $J_1, \ldots, J_I$ in any fashion. Use this partition to define I cost functions

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

[17] We thank Todd Sarver for this observation.

just as in the definition of a temptation representation. Define u_i so that $u_i + \sum_{j \in J_i} v_j = w_i$. Obviously, then, we can write

$$V(x) = \sum_{i=1}^{I} \max_{\beta \in x} [u_i(\beta) - c_i(\beta, x)].$$

Interpreting the I states as equally likely, this looks like a temptation representation where the normative preference, u_i, varies with i. On the other hand, it is not clear what justifies interpreting the u_i's as various possible normative preferences. In our temptation representation, u represents the commitment preference and thus is identified. Note that the inability to identify the u_i's above leads to a more general inability to identify which temptations are relevant in what states since the partition above was arbitrary.

This observation points to another important direction to extend the current model. Our assumption that the normative preference is state independent allows the possibility of identifying at least some aspects of the representation in the sense that these aspects are uniquely determined (up to some transformation). It is not hard to show that the representation is identified in a natural sense if, for example, the u and the various v_j's are *affinely independent* in the sense that these functions (viewed as vectors in $\mathbf{R}^K$ where K is the number of pure outcomes) and the vector of 1's are linearly independent. With such identification, it is possible to consider how changes in preferences correspond to changes in the representation (i.e., analogs to the correspondence between increased willingness to undertake risk and a lower Arrow–Pratt measure of risk aversion). For example, DLR show that one preference has an additive EU 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 comparisons of interest.

Finally, our characterization of behavior which can be explained by temptation is carried out within the set of preferences which have a finite additive EU representation, a set characterized in Theorem 6 in the appendix. While some of the axioms required seem unrelated to issues of temptation, two of the necessary conditions, continuity and independence (see appendix for definitions), 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 her choice from the menu to be determined by her later self with utility function v, where her 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. For 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. Independence implies that such a dieter would be indifferent between this menu and one which adds a randomization between the two. However, 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. Hence the indifference required by independence is not compelling.[18]

Also, there is a sense in which independence implies that the agent's choices satisfy a kind of "independence of irrelevant alternatives." To understand this, note that we represent the agent as if she would face cost function c_i with probability q_i. Subject to the caveats mentioned in footnote 14, suppose we interpret the agent who faces menu x as choosing some β which maximizes $u(\beta) - c_i(\beta, x)$ with probability q_i. Substituting for c_i, this means that the agent maximizes a certain sum of utilities which is independent of x. Hence if β is chosen over α from menu x, β is chosen over α from any menu, a kind of IIA property. This conclusion is driven by the linearity of the representation — this causes the $\max_{\beta \in x} v_j(\beta)$ terms to be irrelevant to the $\max_{\beta \in x} u(\beta) - c_i(\beta, x)$ expression. This linearity comes from independence.

As Noor [2006b] suggests by example, this IIA property is not a compelling assumption for temptation. For a diet–related version of his example, suppose the menu consists only of broccoli and frozen yogurt. Arguably, the latter is not very tempting, so the agent is able to stick to her diet and orders broccoli. However, if the menu consists of broccoli, frozen yogurt, and an ice cream sundae, perhaps the agent is much more significantly tempted to order dessert and opts for the frozen yogurt as a compromise. See also the related criticism of independence in Fudenberg and Levine [2005].

Related to the discussion of guilt two paragraphs above, issues of guilt and its flip side, feeings of "virtuousness," may be important aspects of temptation and pose new modeling challenges. To see the point, we again let b denote broccoli, y frozen yogurt, and i ice cream and assume $\{b\} \succ \{y\} \succ \{i\}$. Suppose the agent knows she will choose y from any menu containing it. Then it seems plausible that $\{y, i\} \succ \{y\} \succ \{b, y\}$. Intuitively, the first preference comes about because the agent can feel virtuous by choosing frozen yogurt over the more fattening ice cream, a feeling which the agent cannot get from choosing yogurt when it is the only option. Similarly, the second preference reflects the agent's guilt from choosing frozen yogurt when broccoli was available, a feeling not

[18]We thank Phil Reny for suggesting this example. The example has a strong resemblance to the "Machina's mom" story in Machina [1989]. The resemblance suggests that the issue is more about having preferences over procedures for decision making, perhaps driven by temptation, than about temptation given otherwise standard preferences, the case we study here.

generated by consuming frozen yogurt when there is no other option. Note that the first of these preferences contradicts our main axiom, DFC, since it implies $\{y, i\} \succ \{y\} \succ \{i\}$. This story also runs contrary to the motivation for our AIC axiom: here, adding i improves the menu $\{y\}$ but does so because it is *not* chosen. While the preference $\{b\} \succ \{y\} \succ \{b, y\}$ is consistent with our general representation, it is not consistent with a simple representation. In particular, with guilt, an agent who succumbs to temptation does not avoid all costs. We suspect that an adequate treatment of these issues requires moving beyond the class of finite additive EU representations.

A Notational Conventions

Throughout the Appendix, we use u, v_j, etc., to denote utility functions as well as the vector giving the payoffs to the pure outcomes associated with the utility function. When interpreted as vectors, they are column vectors. Let K denote the number of pure outcomes, so these are K by 1. We write lotteries as 1 by K row vectors, so $\beta \cdot u = u(\beta)$, etc. Also, $\mathbf{1}$ denotes the K by 1 vector of 1's.

B Existence of Finite Additive EU Representations

It is simpler to work with the following equivalent definition of a finite additive EU representation.

Definition 5 *A* finite additive EU representation *is a pair of finite sets S_1 and S_2 and a state–dependent utility function $U : \Delta(B) \times (S_1 \cup S_2) \to \mathbf{R}$ such that (i) $V(x)$ defined by*

$$V(x) = \sum_{s_1 \in S} \max_{\beta \in x} U(\beta, s) - \sum_{s \in S_2} \max_{\beta \in x} U(\beta, s)$$

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

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

The relevant axioms from DLR are:

Axiom 8 (Weak Order) *$\succ$ is asymmetric and negatively transitive.*

Axiom 9 (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).*

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 $\lambda\beta'(b) + (1 - \lambda)\beta''(b)$.

Axiom 10 (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.

The new axiom which will imply finiteness requires a definition. Given any menu x, let $\mathrm{conv}(x)$ denote its convex hull.

Definition 6 *$x' \subseteq \mathrm{conv}(x)$ is* critical for x *if for all y with $x' \subseteq \mathrm{conv}(y) \subseteq \mathrm{conv}(x)$, we have $y \sim x$.*

Intuitively, a critical subset of x contains all the "relevant" points in x. It is easy to show that the three axioms above imply that the boundary of x is critical for x, so every set has at least one critical subset.

Axiom 11 (Finiteness) *Every menu x has a finite critical subset.*

Theorem 6 *$\succ$ has a finite additive EU representation if and only if it satisfies weak order, continuity, independence, and finiteness.*

Necessity is straightforward. The sufficiency argument follows that of DLR and DLRS by constructing an artificial "state space," S^K, then restricting it to a particular subset. To do this, write $B = \{b_1, \ldots, b_K\}$. Let $S^K = \{s \in \mathbf{R}^K \mid \sum s_i = 0, \sum s_i^2 = 1\}$. In line with our notational conventions, we write elements of S^K as K by 1 column vectors. For any set $x \in X$, let σ_x denote its *support function*. That is, $\sigma_x : S^K \to \mathbf{R}$ is defined by

$$\sigma_x(s) = \max_{\beta \in x} \beta \cdot s.$$

As explained in DLR, our axioms imply that if $\sigma_x = \sigma_{x'}$, then $x \sim x'$.

To prove sufficiency, fix any sphere, say x^*, in the interior of $\Delta(B)$. By finiteness, x^* has a finite critical subset. Let x_c denote such a subset. We claim that we may as well assume x_c is contained in the boundary of x^*. To see this, suppose it is not. For every point in x_c, associate any line through this point. Let $\hat{x}_c$ denote the collection of intersections of these lines with the boundary of x^*. Obviously, $\hat{x}_c$ is finite. Also, it is easy to see that $\mathrm{conv}(x_c) \subseteq \mathrm{conv}(\hat{x}_c)$. In light of this, consider any convex $y \subseteq x^*$ and suppose $\hat{x}_c \subseteq y$. Then

$$x_c \subseteq \mathrm{conv}(x_c) \subseteq \mathrm{conv}(\hat{x}_c) \subseteq y \subseteq x^*.$$

So $y \sim x^*$. Hence $\hat{x}_c$ is a finite critical subset of x^* which is contained in the boundary of x^*. So without loss of generality, we assume x_c is contained in the boundary of x^*.

Since x^* is a sphere, there is a one–to–one mapping, say g, from the boundary of x^* to S^K where $g(\beta)$ is the s such that β is the unique maximizer of $\alpha \cdot s$ over $\alpha \in x$. That is, $g(\beta)$ is the s for which we have an indifference curve tangent to x^* at β. Let

$$S^* = g(x_c) = \{s \in S^K \mid g(\beta) = s \text{ for some } \beta \in x_c\}.$$

Let

$$x = \bigcap_{\beta \in x_c} \{\alpha \in \Delta(B) \mid \alpha \cdot g(\beta) \leq \beta \cdot g(\beta)\}.$$

That is, x is the polytope bounded by the hyperplanes tangent to x^* at the points in x_c.

Lemma 3 *x_c is critical for x.*

Proof. Obviously, $x_c \subset x$. Fix any convex y such that $x_c \subseteq y \subseteq x$. We show that $y \sim x$.

To show this, fix any $\varepsilon > 0$ and let

$$y^\varepsilon = \text{conv}\left(x_c \,\cup\, \left[\bigcap_{\beta \in x_c} \{\alpha \in y \mid \alpha \cdot g(\beta) \leq \beta \cdot g(\beta) - \varepsilon\}\right]\right).$$

Note that $x_c \subseteq y^\varepsilon \subseteq y$. Also, $y^\varepsilon \to y$ as $\varepsilon \downarrow 0$ since $x_c \subseteq y \subseteq x$.

We claim that

Claim 1 *For every $\varepsilon > 0$, there exists $\lambda < 1$ such that*

$$\lambda \text{conv}(x_c) + (1 - \lambda) y^\varepsilon \subseteq x^*.$$

We establish this geometric property shortly. First, note that with this claim, the proof of the lemma can be completed as follows. Fix any $\varepsilon > 0$ and $\lambda \in (0, 1)$ such that $\lambda \text{conv}(x_c) + (1 - \lambda) y^\varepsilon \subseteq x^*$. Because $x_c \subseteq y^\varepsilon$, we have

$$x_c \subseteq \lambda \text{conv}(x_c) + (1 - \lambda) y^\varepsilon \subseteq x^*.$$

Since x_c is critical for x^* and $\lambda \text{conv}(x_c) + (1 - \lambda) y^\varepsilon$ is convex, this implies $\lambda \text{conv}(x_c) + (1 - \lambda) y^\varepsilon \sim x^*$. The fact that x_c is critical for x^* also implies $\text{conv}(x_c) \sim x^*$. Hence independence requires $y^\varepsilon \sim x^*$. Since this is true for all $\varepsilon > 0$, continuity implies $y \sim x^*$.

But this argument also works for the case of $y = x$, so we see that $x \sim x^*$. Hence $y \sim x$, so x_c is critical for x.

Proof of Claim 1. First, note that it is sufficient to prove this for the case of $y = x$ since this makes the set on the left–hand side the largest possible. Next, note that it is then sufficient to show that for every $\varepsilon > 0$, there exists $\lambda < 1$ such that every extreme point of $\lambda \text{conv}(x_c) + (1-\lambda)x^\varepsilon$ is contained in x^*. Since each such extreme point must be a convex combination of extreme points in x_c and x^ε, this implies that a sufficient condition is that there is a $\lambda < 1$ such that for every $\alpha_1 \in x_c$ and $\alpha_2 \in \text{ext}(x^\varepsilon)$, $\lambda\alpha_1 + (1-\lambda)\alpha_2 \in x^*$ where $\text{ext}(\cdot)$ denotes the set of extreme points. Since x^ε is a convex polyhedron, it has finitely many extreme points. Also, x_c is finite. Since there are finitely many α_1 and α_2 to handle, it is sufficient to show that for every $\alpha_1 \in x_c$ and $\alpha_2 \in \text{ext}(x^\varepsilon)$, there is a $\lambda \in (0,1)$ such that $\lambda\alpha_1 + (1-\lambda)\alpha_2 \in x^*$.

Equivalently, we show that for every $\alpha_1 \in x_c$ and $\alpha_2 \in x^\varepsilon$, there exists $\lambda \in (0,1)$ such that $(\lambda\alpha_1 + (1-\lambda)\alpha_2) \cdot s \leq \sigma_{x^*}(s)$ for all $s \in S^K$. That is,

$$(1-\lambda)(\alpha_2 \cdot s - \alpha_1 \cdot s) \leq \sigma_{x^*}(s) - \alpha_1 \cdot s, \quad \forall s \in S^K. \tag{3}$$

Since $\alpha_1 \in x^*$, we have $\sigma_{x^*}(s) \geq \alpha_1 \cdot s$ for all $s \in S^K$. By construction, there is a unique s, say $\hat{s} = g(\alpha_1)$, such that this inequality holds with equality. For all $s \neq \hat{s}$, $\sigma_{x^*}(s) > \alpha_1 \cdot s$. Also, by definition of x^ε, $\alpha_2 \in x^\varepsilon$ implies that $\alpha_2 \cdot \hat{s} \leq \alpha_1 \cdot \hat{s} - \varepsilon$. Hence for any $\lambda \in [0,1]$, equation (3) holds at $s = \hat{s}$. For any $s \neq \hat{s}$, if $\alpha_2 \cdot s \leq \alpha_1 \cdot s$, again, equation (3) holds for all $\lambda \in [0,1]$. Hence we can restrict attention to s such that $\alpha_2 \cdot s > \alpha_1 \cdot s$ and $\sigma_{x^*}(s) > \alpha_1 \cdot s$. Given this restriction, it is clear that if $\alpha_2 \cdot s \leq \sigma_{x^*}(s)$, again, equation (3) holds for all $\lambda \in [0,1]$.

Let $\hat{S} = \{s \in S^K \mid \alpha_2 \cdot s > \sigma_{x^*}(s) > \alpha_1 \cdot s\}$. From the above, it is sufficient to show the existence of a $\lambda \in (0,1)$ satisfying equation (3) for all $s \in \hat{S}$. A sufficient condition for this is that there exists $\lambda \in (0,1)$ such that

$$(1-\lambda)(\sigma_{\Delta(B)}(s) - \alpha_1 \cdot s) \leq \sigma_{x^*}(s) - \alpha_1 \cdot s, \quad \forall s \in \hat{S}.$$

Obviously, $\sigma_{\Delta(B)}(s) - \alpha_1 \cdot s$ is bounded from above. Hence it is sufficient to show that the right–hand side of the inequality is bounded away from zero for $s \in \hat{S}$.

To see that this must hold, suppose there is a sequence $\{s^n\}$ with $s^n \in \hat{S}$ for all n with $\sigma_{x^*}(s^n) - \alpha_1 \cdot s^n \to 0$. Clearly, this implies $s^n \to \hat{s}$. But then

$$\lim_{n\to\infty} \alpha_2 \cdot s^n = \alpha_2 \cdot \hat{s} \leq \sigma_{x^*}(\hat{s}) - \varepsilon = \lim_{n\to\infty} \sigma_{x^*}(s^n) - \varepsilon,$$

implying that we cannot have $s^n \in \hat{S}$ for all n, a contradiction. Hence such a λ must exist. ▮

Lemma 4 *If y is any set with $\sigma_y(s) = \sigma_x(s)$ for all $s \in S^*$, then $y \sim x$.*

Proof: Fix any such y. Without loss of generality, assume y is convex. (Otherwise, we can replace y with its convex hull.) Clearly,

$$y \subseteq \{\beta \mid \beta \cdot s \leq \sigma_x(s) \ \ \forall s \in S^*\}$$

since otherwise y would contain points giving it a higher value of the support function for some $s \in S^*$. But the set on the right–hand side is x, so $y \subseteq x$. Obviously, then if $x_c \subseteq y$, the fact that x_c is critical for x implies $y \sim x$.

So suppose $x_c \not\subseteq y$. As noted, we must have $y \subseteq x$. So let $y_\lambda = \lambda x + (1-\lambda)y$. Obviously, y_λ converges to x as $\lambda \to 1$. For each $\beta \in x_c$, there is a face of the polyhedron x such that β is in the (relative) interior of the face. Also, y must intersect the face of the polyhedron and so y_λ must intersect the face. As λ increases, the intersection of y_λ with the face enlarges as it is pulled out toward the boundaries of the face. Clearly, for λ sufficiently large, β will be contained in the intersection of y_λ with the face of x which contains β. Take any λ larger than the biggest such λ over the finitely many $\beta \in x_c$. Then $x_c \subseteq y_\lambda \subseteq x$. Since x_c is critical for x, this implies $\lambda x + (1-\lambda)y \sim x$. By independence, then, $y \sim x$. ▌

Lemma 5 *For any y and $\hat{y}$ such that $\sigma_y(s) = \sigma_{\hat{y}}(s)$ for all $s \in S^*$, we have $y \sim \hat{y}$.*

Proof. Fix any such y and $\hat{y}$. For any $\lambda \in [0,1)$, define $u_\lambda : S^* \to \mathbf{R}$ by

$$u_\lambda(s) = \frac{\sigma_x(s) - \lambda\sigma_y(s)}{1-\lambda}.$$

Because $\sigma_y(s) = \sigma_{\hat{y}}(s)$ for all $s \in S^*$, it would be equivalent to use $\sigma_{\hat{y}}$ instead of σ_y. Let

$$z_\lambda = \{\beta \in \Delta(B) \mid \beta \cdot s \leq u_\lambda(s), \ \ \forall s \in S^*\}.$$

Obviously, $\lambda\sigma_y(s)+(1-\lambda)u_\lambda(s) = \sigma_x(s)$ for all $s \in S^*$. This implies that for all $\lambda \in (0,1)$, $\lambda y + (1-\lambda)z_\lambda \subseteq x$. To see this, note that for any $\alpha \in y$ and $\beta \in z_\lambda$,

$$\lambda\alpha \cdot s + (1-\lambda)\beta \cdot s \leq \lambda\sigma_y(s) + (1-\lambda)u_\lambda(s) = \sigma_x(s), \ \ \forall s \in S^*.$$

But $x = \cap_{s \in S^*}\{\gamma \mid \gamma \cdot s \leq \sigma_x(s)\}$, so $\lambda\alpha + (1-\lambda)\beta \in x$.

Note also that $u_\lambda(s) \to \sigma_x(s)$ as $\lambda \downarrow 0$. We claim that this implies that there is a $\lambda \in (0,1)$ such that for every $s \in S^*$, there exists $\beta \in z_\lambda$ with $\beta \cdot s = u_\lambda(s)$. To see this, suppose it is not true. Then for all $\lambda \in (0,1)$, there exists $\hat{s}_\lambda \in S^*$ such that for all $\beta \in z_\lambda$, $\beta \cdot \hat{s}_\lambda < u_\lambda(\hat{s}_\lambda)$ so

$$\bigcap_{s \in S^* \setminus \{\hat{s}_\lambda\}} \{\beta \mid \beta \cdot s \leq u_\lambda(s)\} = \bigcap_{s \in S^*} \{\beta \mid \beta \cdot s \leq u_\lambda(s)\}.$$

Because S^* is finite, this implies that there exists $\hat{s} \in S^*$, a sequence $\{\lambda_n\}$ with $\lambda_n \in (0,1)$ for all n, $\lambda_n \to 0$ such that for all n,

$$\bigcap_{s \in S^* \setminus \{\hat{s}\}} \{\beta \mid \beta \cdot s \leq u_{\lambda_n}(s)\} = \bigcap_{s \in S^*} \{\beta \mid \beta \cdot s \leq u_{\lambda_n}(s)\}.$$

But $u_{\lambda_n} \to \sigma_x$ as $n \to \infty$. Hence the limit as $n \to \infty$ of the right–hand side, namely x, cannot equal the limit of the left–hand side, a contradiction.

Hence there is a $\lambda \in (0,1)$ such that for every $s \in S^*$, there is a $\beta \in z_\lambda$ with $\beta \cdot s = u_\lambda(s)$. Choose such a λ and let $u = u_\lambda$ and $z = z_\lambda$. Obviously, for every $s \in S^*$, there is $\alpha \in y$ with $\beta \cdot s = \sigma_y(s)$. Hence given our choice of λ, for every $s \in S^*$, there is $\gamma \in \lambda y + (1-\lambda) z$ such that $\gamma \cdot s = \lambda \sigma_y(s) + (1-\lambda) u(s) = \sigma_x(s)$. Hence $\sigma_{\lambda y + (1-\lambda) z}(s) = \sigma_x(s)$ for all $s \in S^*$. Hence Lemma 4 implies $\lambda y + (1-\lambda) z \sim x$. The symmetric argument with $\hat{y}$ replacing y implies $\lambda \hat{y} + (1-\lambda) z \sim x$. So $\lambda y + (1-\lambda) z \sim \lambda \hat{y} + (1-\lambda) z$. By independence, then, $y \sim \hat{y}$. ▮

DLR show that weak order, continuity, and independence imply the existence of a function $V : X \to \mathbf{R}$ which represents the preference and is affine in the sense that $V(\lambda x + (1-\lambda) y) = \lambda V(x) + (1-\lambda) V(y)$. Fix such a V. Let $\mathcal{U} = \{(\sigma_x(s))_{s \in S^*} \mid x \in X\} \subset \mathbf{R}^M$ where M is the cardinality of S^*. Let $\sigma | S^*$ denote the restriction of σ to S^*. Define a function $W : \mathcal{U} \to \mathbf{R}$ by $W(U) = V(x)$ for any x such that $\sigma_x | S^* = U$. From Lemma 5, we see that if $\sigma_x | S^* = \sigma_{x'} | S^*$, then $x \sim x'$ so $V(x) = V(x')$. Hence W is well–defined. It is easy to see that W is affine and continuous and that $\mathcal{U}$ is closed, convex, and contains the 0 vector. It is easy to show that W has a well–defined extension to a continuous, linear function on the linear span of $\mathcal{U}$. Since $\mathcal{U}$ is finite dimensional, W has an extension to a continuous linear functional on $\mathbf{R}^M$. (See Lemma 6.13 in Aliprantis and Border [1999], for example.) Since a linear function on a finite dimensional space has a representation by means of a matrix, we can write

$$W(U) = \sum_{s \in S^*} c_s U_s$$

where the c_s's are constants and $U = (U_s)_{s \in S^*}$. Hence

$$V(x) = W((\sigma_x(s))_{s \in S^*}) = \sum_{s \in S^*} c_s \max_{\beta \in x} \beta \cdot s.$$

Hence we have a finite additive EU representation. ▮

C Proof of Theorem 3

The following lemma is critical.[19]

Lemma 6 *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 KI 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

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

Write $p = (p_1, \ldots, p_I)$ where each p_i is a 1 by K 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,

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

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

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 non–negative 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 $(p_{1i}, \ldots, p_{Ki})$. Without loss of generality, we can assume that $|p_{ki}| \leq 1/K$ for all k and i. (Otherwise we could divide both sides of the inequality above by $K \max_{k,i} |p_{ki}|$ and redefine p_i to have this property.) Let β denote the probability distribution $(1/K, \ldots, 1/K)$. For each i, let $\alpha_i = p_i + \beta$. Note that $\alpha_{ki} = p_{ki} + 1/K$ 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 3. 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 $i = 1, \ldots, I'$, let α_i denote a maximizer of $u(\beta) + \sum_{j \in J_i} v_j(\beta)$ over $\beta \in x$. Then

$$\begin{aligned} V_w(x) &= \textstyle\sum_{i=1}^{I'} q_i[u(\alpha_i) + \sum_{j \in J_i} v_j(\alpha_i)] - \sum_{i=1}^{I'} q_i \sum_{j \in J_i} \max_{\beta \in x} v_j(\beta) \\ &\qquad + \textstyle\sum_{i=I'+1}^{I} \max_{\beta \in x}[-c_i(\beta, x)] \\ &\leq \textstyle\sum_{i=1}^{I'} q_i[u(\alpha_i) + \sum_{j \in J_i} v_j(\alpha_i)] - \sum_i q_i \sum_{j \in J_i} v_j(\alpha_i) \\ &= \textstyle\sum_{i=1}^{I'} q_i u(\alpha_i) \\ &\leq \max_{\beta \in x} u(\beta) \end{aligned}$$

where the first inequality uses $c_i(\beta, x) \geq 0$ for all i, β, and x and the last one uses $q_i > 0$ and $\sum_{i=1}^{I'} q_i = 1$. Hence DFC must hold.

For sufficiency, let V denote a finite additive EU representation of $\succ$. By Lemma 6, we can write this as

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

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 M 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 $\{1, \ldots, M\}$. 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,

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

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. ▮

D Proof of Theorem 2

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.

Recall that

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

Lemma 7 *If $\succ$ has a temptation representation, then it satisfies AIC.*

Fix $\succ$ and a temptation representation, V_T. Let β be an approximate improvement for x. Fix any $x' \subseteq x$ and $\alpha \in B(x')$ such that $\{\alpha\} \succ \{\beta\}$. (If no such x, β, x', and α exist, AIC 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 $x' \subseteq x$,

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

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

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

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

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

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

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

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 of the w_i's and v_j's 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

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

for an arbitrary constant A. Let w_1 denote a constant function equal to A 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 $V(x) = \max_{\beta \in x} w_1(\beta) + A$ for an arbitrary constant A. Let v_1 be a constant function equal to A 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 $\{\alpha\} \sim x$.

Lemma 8 *Suppose $\succ$ satisfies AIC 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 $w_i(\gamma_j) < w_i(\beta)$. 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,

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

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

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 AIC implies that $x \cup \{\beta\}$ cannot be temptation–free, a contradiction. ▌

To complete the proof of Theorem 2, we use the following result from Rockafellar [1970] (Theorem 22.2, pages 198–199):

Lemma 9 *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 AIC. We use Lemma 9 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 9 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 9, 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 $\lambda_{ik} < 1/K$ for all i and k. (Recall that there are K pure outcomes.) Otherwise, we can divide through all equations by $2K \max_{i,k} |\lambda_{ik}|$ 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 $\alpha_i(k) = (1/K) - \lambda_{ik}$. Because $\lambda_{ik} < 1/K$ 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 $(1/K, \ldots, 1/K)$, 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

$$u(\beta) + \bar{\mu} \geq \max_i u(\alpha_i) \tag{4}$$

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

$$\sum_j v_j(\beta) + \sum_j \mu_j \geq \sum_j \max_i v_j(\alpha_i). \tag{5}$$

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 (4) and (5) 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

$$\sum_i w_i(\alpha_i) - \sum_j \max_i v_j(\alpha_i) \geq u(\beta) + \sum_j v_j(\beta) + \bar{\mu} + \sum_j \mu_j - \sum_j \max_i v_j(\alpha_i) \geq \max_i u(\alpha_i). \quad (6)$$

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 (6) implies that the weak inequalities in equations (4) and (5) 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 8.

Hence there must exist such a_i, b_{ij}, and c_i. It is easy to use the proof of Theorem 3 to complete the construction of a temptation representation. ▮

E 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

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

Hence

$$\begin{aligned} V(x \cup y) &\leq \max\Big\{\max_{\beta\in x} w_1(\beta), \max_{\beta\in y} w_1(\beta)\Big\} \\ &\quad - \max\Big\{\textstyle\sum_j \max_{\beta\in x} v_j(\beta), \sum_j \max_{\beta\in y} v_j(\beta)\Big\} \\ &\leq \max\Big\{\max_{\beta\in x} w_1(\beta) - \textstyle\sum_j \max_{\beta\in x} v_j(\beta), \\ &\qquad\qquad \max_{\beta\in y} w_1(\beta) - \textstyle\sum_j \max_{\beta\in y} v_j(\beta)\Big\} \\ &= \max\{V(x), V(y)\} = V(x). \end{aligned}$$

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. ▌

F Proof of Theorem 5

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 6, 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 3 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 eliminate every 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] Aliprantis, C., and K. Border, *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Berlin: Springer–Verlag, 1999.

[2] Amador, M., I. Werning, and G.-M. Angeletos, "Commitment vs. Flexibility," *Econometrica*, **74**, March 2006, 365–396.

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

[4] Dekel, E., B. Lipman, A. Rustichini, and T. Sarver, "Representing Preferences with a Unique Subjective State Space: Corrigendum," working paper, December 2005.

[5] Ergin, H., and T. Sarver, "A Unique Costly Contemplation Representation," working paper, December 2005.

[6] Fudenberg, D., and D. Levine, "A Dual Self Model of Impulse Control," working paper, August 2005.

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

[8] Gul, F., and W. Pesendorfer, "The Simple Theory of Temptation and Self–Control," working paper, 2005.

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

[10] Kopylov, I., "Temptations in General Settings," working paper, 2005.

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

[12] 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.

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

[14] Noor, J., "Temptation, Welfare, and Revealed Preference," Boston University working paper, 2006a.

[15] Noor, J., "Menu–Dependent Self–Control," Boston University working paper, 2006b.

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

[17] Sarver, T., "Anticipating Regret: Why Fewer Options May Be Better," working paper, November 2005.

[18] 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.