



Ulysses revisited: how an expressed preference for external control reveals the limit of willpower

Meng-Yu Liang¹ · Simon Grant²

Accepted: 6 April 2025 / Published online: 9 May 2025
© The Author(s) 2025

Abstract

We use the revealed preference approach to identify a preference for external control for a decision maker facing a binding willpower constraint in the absence of such external control. We revisit the classic story of Ulysses and the Sirens to demonstrate that the choice behavior (over menus) exhibited by Ulysses can neither be accommodated by Gul and Pesendorfer (2001)’s model of costly self-control temptation preferences nor by Noor and Takeoka (2010)’s extension to an increasing (that is, “uphill”) self-control cost.

Keywords Dynamic Inconsistency · Temptation and Self-Control · Willpower · Revealed preferences

1 Introduction

Often with tractability and theoretical parsimony in mind, the standard approach to modeling economic actors is to envision them as (mindless) hyper-rational decision-makers, who can discern and then implement the optimal course of action: unencumbered by any conflicting emotions; impervious to temptation; and, not bound by any limitation on willpower. However, as Loewenstein (2000b) notes:

“[w]hen visceral factors propel behavior in directions that are not commensurate with self-interest, ... that is, when strong emotions, drives, and somatic

Liang would like to acknowledge financial support from Taiwan’s National Science and Technology Council under grant 109-2410-H-001-041.

✉ Meng-Yu Liang
myliang@econ.sinica.edu.tw

Simon Grant
simon.grant@anu.edu.au

¹ Institute of Economics, Academia Sinica, Taipei, Taiwan

² Australian National University, Canberra, Australia

sensations are in play, decision making models that do not incorporate will-power will be fatally incomplete. They predict how people wish they could behave rather than how they actually behave.”

During the past two decades, Loewenstein’s warnings have been echoed by an emerging literature that explicitly addresses issues of temptation and costly self-control. Gul and Pesendorfer (2001) (hereafter, GP) were the first to provide a characterization for temptation preferences by considering the preferences of a decision-maker (hereafter, DM) over menus of alternatives in which the evaluation of the menus available to the DM takes place in an initial period when the DM is in a “cold” (emotional) state. Then in a subsequent period, the DM will choose an item from the menu to which she committed herself in the initial period. Depending on what items are available in that menu, she may be in a “hot” state and so experience “temptation” cravings that require her to incur a self-control cost if she does not choose that menu’s most tempting item.

By expanding the preference domain, GP can reconcile the time-inconsistency problems that arise in other approaches to modeling an individual with conflicting desires across time (and across emotional states). If a commitment device is available that allows the DM in the initial period to commit to a single-item menu, then she has constrained her choice in the subsequent period to adhere to how in the initial period she wished she would behave in the subsequent period. However, this perfect resolution entails a substantial limitation. When a commitment device removes a tempting alternative from a menu, the visceral factor must disappear too. Recalling GP’s motivating story about a DM who in the morning is deciding which restaurant to commit to go to for lunch, the DM’s craving for hamburger will never be present if hamburger is not on the menu.

To illustrate what important aspect of the struggle between temptation and self-control GP’s model misses, we consider the story from Greek mythology of Ulysses and the Sirens. The Sirens were infamous nymphs who had the power of charming by their song all mariners to such an extent that those mariners felt compelled to cast themselves into the sea and drown. To avert such a disaster, while still allowing Ulysses to experience the enchanting song of the Sirens, Circe suggested Ulysses fill the ears of his crew with wax and to have them tie him to the mast with strict instructions not to untie him until their ship was safely out of earshot of the Sirens.

Notice that there are two distinct commitment devices in the story of Ulysses: ‘have ears filled with wax’ and ‘be tied to the mast’. The former may be viewed as a *hard* commitment device since it enables a DM (in this case, each member of Ulysses’ crew) to comply with the wish behavior, that is, returning home, by *removing* from the choice menu any item that elicits a craving for listening to the music. This is the type of commitment covered in GP’s model. The latter device, however, commits the DM (in this case, Ulysses) to returning home while still bearing the cost of craving for more music. This is the additional commitment device we introduce in this paper. We refer to this as an *external control* commitment device. As we shall see in the next section, such a commitment device is not required if the DM is

not subject to a willpower constraint that limits the self-control cost she is able to bear by not selecting the most tempting alternative in a menu. With the distinction between (self-control) commitment and external control, some of the cases that have been considered as commitment devices may be better understood as external control. For example, in (Bryan et al. (2010), section 2.3.4), the example of taking only a fixed amount of cash when heading out to party for the night is given as a commitment device. Compared to a hard commitment such as not going to a party, this example also excludes the possibility of overspending, while allowing the person to *experience* the temptation of the party as well as incurring any (utility) cost of self-control that may arise from not being able to overspend.¹

Our model, like GP's, maintains a tight correspondence between the behavioral and welfare components of economic models. Hence, unlike the literature on quasi-hyperbolic discounting, we do not suffer the problem of the dual self approach with regard to whose welfare we should maximize (the welfare of the 'planner' or that of the 'doer' whose discount factor is lower than that of the planner) when deciding public policy. We follow (Bernheim and Rangel 2007)'s contention that "revealed preference is an attractive political principle because it guards against abuse." They elaborate by noting:

"There is widespread agreement that normative criteria should respect the principle of individual sovereignty, which holds that notions of good and bad for society should be rooted in the notions of good and bad held by the affected individuals. This principle instructs policy analysts to act as each individual's proxy when comparing alternative policies. It precludes the analyst from imposing his or her own value judgments."

In Ulysses's story, the preference for coercion over discretion in a 'hot' state is a reflection of the DM's limited willpower in such an emotional state. We shall see in the basic extension of GP's and Liang et al. (2020)'s models that we present in section 3, in which the DM is certain in the initial period which item in a given menu will be optimal in the subsequent period, that external-control dominates self discretion. For example in Liang et al. (2020)'s model of costly self-control with limited willpower, self-indulgent behavior by the DM in the subsequent period can be a legitimate rational choice from the doer's perspective. The DM's sovereignty on whether to commit in the initial (cold state) period to an external control device reflects her degree of respect for the welfare of a 'doer' in the subsequent (hot state) period. Hence, she would prefer coercion in a 'hot' state *only if* the payoff structure is similar to the example of Ulysses and the Sirens. For instance, some children exhibit a preference for parental controls that impose time limits on their video game usage, whereas others do not. Similarly, some husbands may favor spousal management of their financial resources, as evidenced by research (see Ashraf 2009), due to concerns that they might engage in excessive spending if they had direct access to their income.

¹ We thank an anonymous referee for suggesting this example.

1.1 Related literature

Models of time inconsistency and the demand for imposing constraints on the DM's future behavior has accumulated a vast literature in economics. One line of the formulations is by changing-tastes, the DM uses different preferences to govern her future consumption; most notable are the Strotz representation and (so-called) quasi-hyperbolic discounting. Strotz (1955), Schelling (1984), Laibson (1996), O'donoghue and Rabin (1999). Another line involves the doer-planner (or dual-self) models by Thaler and Shefrin (1981), Shefrin and Thaler (1988), Bernheim and Rangel (2004), Fudenberg and Levine (2006). They assume that there is a conflict in preferences between the long-run planner and a sequence of myopic doers. A third approach considers temptation preferences and costly self-control (e.g., Gul and Pesendorfer (2004), Noor and Takeoka (2010), Liang et al. (2020)). All three approaches yield very similar positive implications and can be difficult to distinguish (see, e.g., Dekel and Lipman (2012)), on the relationship between random Strotz representations and the GP representation.)

The novelty of this paper is to introduce the option of external control for the selection of an item from a menu in the period subsequent to the initial period in which the menu was chosen.

2 Revisiting Ulysses and the Sirens

To translate the story of Ulysses and the Sirens into GP's framework, first consider the "prizes" (that is, final outcomes) available to Ulysses that depend on the actions he takes. Each prize is characterized by two attributes. The first attribute is one of three (final) locations: *Aeaea* (where Ulysses and his crew have been staying for the past year with Circe), *the bottom of the sea* (drowned), and *home* (that is, Ithaca). We shall denote these three locations by a , b , and h , respectively. The second attribute is whether or not Ulysses experiences the music of the Sirens. We shall denote by m (respectively, 0) the case where he experiences (respectively does not experience) the music of the Sirens. So ℓ_e denotes a generic prize where ℓ is a location in $\{a, b, h\}$ and e is an experience in $\{m, 0\}$.

In the initial period, Ulysses chooses among four possible courses of action that will both determine what alternatives are available in the subsequent period's menu as well as the nature of the control (self or external) to which he has committed himself in making a selection from that menu. First he can choose 'to stay' at Circe's island *Aeaea* or '(to have the ears of his crew filled with wax and)to sail away'.² If he chooses the former then the menu he faces in the subsequent period is the single-item menu $A_1 = \{a_0\}$.

There are three different courses of action he can take should he decide to sail away. The first is to have his own ears filled with wax. This is a commitment device that ensures in the subsequent period he will not be in a 'hot' emotional state as

² To simplify the exposition we have excluded the possibility of Ulysses sailing away with a crew whose ears were not filled with wax.

he will not hear the Siren's song. So he will face the menu $A_2 = \{b_0, h_0\}$. On the other hand, if he chooses not to have his ears filled with wax, then in the subsequent period he will hear the Siren's music putting him in a 'hot' emotional state. His final two options are thus whether or not to have his crew bind him to the mast. If he chooses not to be bound to the mast then the menu he will face in the subsequent period is $A_3 = \{b_m, h_m\}$.

One might be tempted to conclude that if Ulysses instructs his crew to bind him to the mast then he will face the (single-item) menu $A_4 = \{h_m\}$. But this misses the point that although an external control device may prevent Ulysses from selecting b_m , as he is in the hot state, he still craves to hear more music, even if he is compelled to choose h_m . Hence, he is still subject to the *disutility* incurred by not satisfying that craving by not having chosen b_m . That is, if he chooses to be bound to the mast then he also faces the menu $A_3 = \{b_m, h_m\}$ but with an external control that ensures the alternative h_m is selected. We denote this menu with such an external control by $A_3(h_m)$.

The choices that Ulysses faces in the initial period along with the different menus in the subsequent period that result are summarized in the decision tree in figure 1.

The above exposition has demonstrated the need to expand the preference domain to distinguish a menu $A(x)$ (with x being an element of A), that denotes the menu A with an external control ensuring the alternative x is chosen in the subsequent period, from the single-item (commitment) menu $\{x\}$.

To see why the choice over menus exhibited by Ulysses cannot be accommodated by Noor and Takeoka (2010)'s increasing uphill self-control cost model extended to include menus subject to external control, first recall in their model, preferences over menus of lotteries (not subject to external control) admit a representation in which a menu A is assigned a utility

$$U(A) := \max_{x \in A} u(x) - f\left(\max_{y \in A} v(y) - v(x)\right), \quad (1)$$

where A is a (compact) set of lotteries with generic elements denoted by x and y , u and v are von Neumann-Morgenstern utility functions over lotteries that describe the individual's normative and temptation rankings, respectively, and $f(\cdot)$ is an increasing convex function measuring the self-control cost incurred by not choosing the most tempting alternative in a menu.³

To extend this to the menu $A(x)$ (a menu subject to external control that ensures the item x from A will be chosen), we naturally set

$$U(A(x)) := u(x) - f\left(\max_{y \in A} v(y) - v(x)\right). \quad (2)$$

From the story we know the normative utilities for the prizes should satisfy⁴:

³ Recall (Gul and Pesendorfer 2001)'s model is the special case for which $f(v) = v$.

⁴ As the outcome a_m is not available its utility values are not relevant.

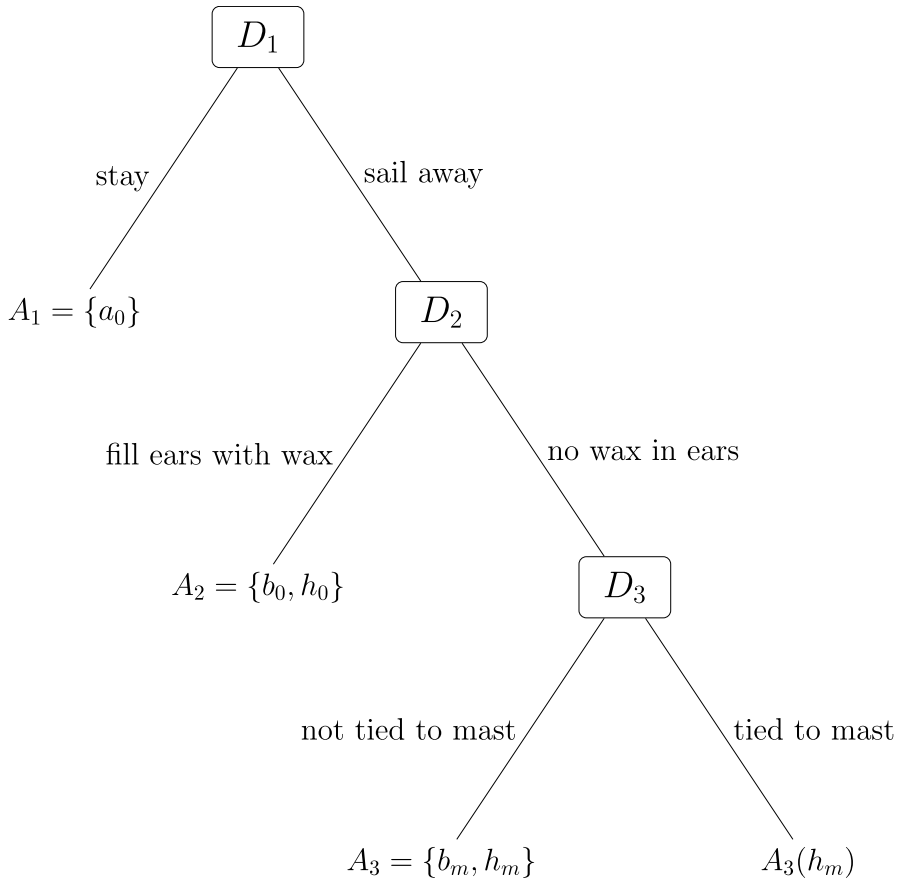


Fig. 1 Decision tree for Ulysses in initial (cold) state

$$u(h_m) > u(h_0) > u(a_0) > u(b_m) > u(b_0).$$

Furthermore, since Ulysses (correctly) anticipates he will choose b_m from the menu $A_3 = \{b_m, h_m\}$ we have $U(A_3) = u(b_m)$, that is,

$$v(b_m) > v(h_m) \quad \text{and} \quad u(b_m) > u(h_m) - f(v(b_m) - v(h_m)).$$

But then that means

$$U(A_3(h_m)) = u(h_m) - f(v(b_m) - v(h_m)) < u(b_m) = U(A_3),$$

which is *inconsistent* with Ulysses instructing his crew to tie him to the mast.

Since we cannot accommodate Ulysses' choices in the initial cold state with Noor and Takeoka (2010)'s model we shall instead employ the *costly self-control with limited willpower* model introduced by Liang et al. (2020) (hereafter, LGH) to represent preferences over menus not subject to external control. This representation

corresponds to GP's model of costly self-control preferences with the addition of a willpower constraint that requires one extra parameter, w , corresponding to the quantity of willpower (measured in units of temptation utility) available to the DM in the subsequent period if she has *not* committed herself to external control. The utility assigned to a menu A for which the DM has not committed herself to external control is given by:

$$U(A) := \max_{x \in A} [u(x) + v(x)] - \max_{y \in A} v(y)$$

$$\text{s.t. } v(x) \geq \max_{y \in A} v(y) - w$$

For preferences over menus subject to external control, however, the willpower constraint is no longer relevant. So we employ a natural extension of the GP model by setting:

$$U(A(x)) := u(x) + v(x) - \max_{y \in A} v(y).$$

Now suppose,

$$v(a_0) \geq v(b_0) \quad \text{and} \quad u(h_m) - u(h_0) > v(b_m) - v(h_m) > w.$$

Intuitively, the first inequality requires that the temptation utility of staying at Aeaea (and hence not experiencing the alluring song of the Sirens) is no less than the temptation utility of jumping into the sea after the Sirens when having not experienced their song. The second and third inequalities require that the difference in normative utilities between going home having heard the music and going home having not heard the music is greater than the difference in temptation utilities between jumping into the sea after the Sirens when having heard their song and simply returning home when having heard their song, and moreover, that latter difference in temptation utilities is greater than Ulysses' stock of willpower.

With these three inequalities, along with the other inequalities stated above, it follows that:

$$U(A_3(h_m)) = u(h_m) - (v(b_m) - v(h_m)) > u(h_0)$$

$$= U(A_2) > u(a_0) = U(A_1) > u(b_m) = U(A_3).$$

Since the commitment menu $\{h_m\}$ is not available to Ulysses, his optimal course of action, reflecting what he likes best, leads to him selecting the menu with external control $A_3(h_m)$. That is, in the initial cold state, Ulysses will order his crew to fill their ears with wax and set sail from Aeaea. He will not put any wax in his own ears but instead will instruct his crew to tie him to the mast with strict orders not to release him until after the ship is out of earshot of the Sirens.

Without these two devices being available, a far-sighted Ulysses would choose to stay at Aeaea. Since $U(A_1) = u(a_0) > u(b_m) = U(A_3)$, he maximizes his long run self's utility by doing so.

Employing just GP's preference domain, in the initial cold state the only valid commitment device that restricts him to choosing a particular item in the

subsequent period is the commitment menu $A_1 = \{h_0\}$. Hence, we can only observe the choice h_0 reflecting his preference $U(\{h_0\}) > U(\{h_m, b_m\})$. A myopic Ulysses, who chooses to sail without an external commitment device, will end up in the subsequent period choosing b_m from $A_3 = \{h_m, b_m\}$. Hence, without the ‘tie to the mast’ action which cannot be encoded in the GP domain, the only available data from observed choices are a_0 (if no wax is available and Ulysses is not myopic), h_0 (if everyone’s ears are filled with wax, including those of Ulysses) and b_m (if no wax is available and Ulysses is myopic). The truly optimal choice of h_m over b_m from the menu $\{h_m, b_m\}$ will never occur in the data.

However, if we expand the domain to include the menu subject to external control that results from the action of having the crew tie Ulysses to the mast (as suggested by the clever Circe), namely the menu $A_3(h_m)$, then the preference for h_m over b_m , reflecting the preference embodied in the inequality

$$u(h_m) - (v(b_m) - v(h_m)) > u(b_m),$$

can be revealed from the choice data.

Therefore, from Ulysses’s long-run self’s perspective (that is, the perspective Ulysses has in the initial cold state), the reason that Ulysses would choose in the subsequent period the option b_m from the menu $\{h_m, b_m\}$ (a menu *without external control*) is due to the willpower constraint he faces in that period rather than any preference he might have in that period. That is, he would choose b_m because he is *incapable of not choosing it* rather than because he does not want to choose h_m .

3 Modeling menus subject to external control

We consider a two-period decision problem similar to the setting in GP. Let (Z, d) be a compact metric space, where Z is the set of (final) prizes (or consequences) and d is a metric defined on Z . Let $\Delta(Z)$ denote the set of all lotteries defined on Z , with generic elements x, y, a, b , et cetera. That is, $\Delta(Z)$ may be taken to be the set of all probability measures on $\mathcal{B}_Z$, the Borel σ -algebra of subsets of Z . We endow $\Delta(Z)$ with the weak topology generated by some metric d_p . As is standard, for any pair of lotteries x, y in $\Delta(Z)$ and any α in $[0, 1]$, let $\alpha x + (1 - \alpha)y$ denote the lottery that assigns to each subset of prizes $B \in \mathcal{B}_Z$, the probability $\alpha x(B) + (1 - \alpha)y(B)$.

Let $\mathcal{A}$ denote the set of menus which we take to be the set of all compact subsets of $\Delta(Z)$ with generic elements A, B, C . We endow $\mathcal{A}$ with the (Hausdorff) topology generated by the metric

$$d_h(A, B) = \max \left\{ \max_{x \in A} \min_{y \in B} d_p(x, y), \max_{x \in B} \min_{y \in A} d_p(x, y) \right\}.$$

For any pair of menus A, B in $\mathcal{A}$ and any α in $[0, 1]$, let $\alpha A + (1 - \alpha)B$ denote the menu in $\mathcal{A}$ given by $\{\alpha x + (1 - \alpha)y : x \in A, y \in B\}$.

For each menu A in $\mathcal{A}$ and each lottery x in A , let $A(x)$ denote the menu subject to *external control* in which the DM in the subsequent period will be compelled to select x from A . Let $\mathcal{A}^*$ denote the set of menus subject to external control. For any pair of menus $A(x)$ and $B(y)$ in $\mathcal{A}^*$ and any α in $[0, 1]$, let $\alpha A(x) + (1 - \alpha)B(y)$ denote the menu $C(\alpha x + (1 - \alpha)y)$ in $\mathcal{A}^*$ in which $C = \alpha A + (1 - \alpha)B$.

The preferences $\succeq$ of the decision-maker (hereafter, DM) are defined on $\mathcal{A} \cup \mathcal{A}^*$.

As is standard, $>$ (respectively, $\sim$) denotes the asymmetric (respectively, symmetric) parts of $\succeq$.

Following the approach taken in LGH, fixing a lottery $x \in \Delta(Z)$, we take the set of tempting alternatives for which the DM fails to exert self-control as ones in which the DM is unable or unwilling to exert self-control in a two-element menu comprising x and that alternative. That is, an alternative y is deemed a *tempting alternative to x for which the DM fails to exert self-control*, if, despite strictly preferring, according to her normative preferences, lottery x to lottery y , she is indifferent in the initial cold state between $\{y\}$ and the menu $\{x, y\}$. We interpret the latter indifference as reflecting her (rational) anticipation that if in the subsequent hot state she faces a choice from the menu $\{x, y\}$ she will be unable or unwilling to exert (sufficient) self-control to choose x . Formally,

$$T(x) := \{y \in \Delta(Z) : \{x\} > \{x, y\} \sim \{y\}\}.$$

Correspondingly, we define the set of *tempting alternatives to x for which the DM can exert costly self-control* as being,

$$S(x) := \{y \in \Delta(Z) : \{x\} > \{x, y\} > \{y\}\}.$$

Now for any lottery y in $T(x)$, there are two distinct reasons for why the DM might anticipate she would not choose x from the menu $\{x, y\}$: (i) the alternative y is at least as good a “compromise” alternative as x , or (ii) despite x being the better compromise candidate, the individual has insufficient willpower to resist choosing y . In the latter case, by considering another alternative $(1 - \alpha)x + \alpha y$ that is formed by taking a convex combination of x and y sufficiently close to x , that is, α is sufficiently close to 0, the DM will be able (and willing) to exert self-control when making a choice from the menu $\{x, (1 - \alpha)x + \alpha y\}$. Hence we divide $T(x)$ into two sets, $I(x)$ and $D(x)$, where

$$I(x) := \{y \in T(x) : (1 - \alpha)x + \alpha y \in S(x), \text{ for some } \alpha \in (0, 1)\}, \text{ and}$$

$$D(x) := \{y \in T(x) : (1 - \alpha)x + \alpha y \notin S(x), \text{ for all } \alpha \in (0, 1)\}.$$

We view $I(x)$ as containing those alternatives in $T(x)$ for which the DM *would like to but has insufficient willpower to exert the self-control necessary to resist the tempting alternative*, while $D(x)$ contains those alternatives for which the DM prefers to indulge herself.

LGH introduced the following class of preferences over menus (not subject to external control) that generalizes GP’s temptation preferences by adding a willpower constraint that embodies the idea that there is a menu-invariant upper-bound on the amount of self-control cost a DM is able to bear arising from not

selecting the most tempting alternative from the menu to which she committed herself in the initial period.

Definition 1 Fix a preference relation $\succsim$ for which $I(a) \neq \emptyset$ for some $a \in \Delta(Z)$. The restriction of $\succsim$ to $\mathcal{A}$ is a member of the class of *limited (willpower) self-control* preference relations characterized by the triple $\langle u, v, w \rangle$, where u and v are continuous linear non-constant functions defined on $\Delta(Z)$ with v not an affine transformation of u , except for when $v(\cdot) = -\alpha u(\cdot) + \beta$ for some $\alpha \in (0, 1)$ and $\beta \in \mathbb{R}$, and $w \geq 0$, if the restriction can be represented by the functional given by:

$$U(A) := \max_{x \in A} [u(x) + v(x)] - \max_{y \in A} v(y), \quad (3)$$

$$\text{s.t. } \max_{y \in A} v(y) - v(x) \leq w,$$

We refer to such a functional $U(\cdot)$ as a *limited self-control* functional.⁵

As our rendition of the story of Ulysses and the Sirens demonstrated, in addition to menus not subject to external control, a DM may have available to her an external commitment device that allows her not only to choose a menu in the initial (cold state) period but also to commit herself to selecting a particular alternative from that menu in the subsequent (hot state) period. That is, the external commitment device essentially removes the willpower constraint she faces with a menu not subject to external control. To extend the DM's preferences in the initial period to menus subject to external control, we propose this natural extension of the GP class of temptation with costly self-control preferences.

Definition 2 Fix a preference relation $\succsim$. The restriction of $\succsim$ to $\mathcal{A}^*$ is a member of the class of *no-limit external-control* preference relations characterized by the pair $\langle u, v \rangle$, where u and v are linear functions defined on $\Delta(Z)$, if the restriction can be represented by the functional given by:

$$U(A(x)) = u(x) + v(x) - \max_{y \in A} v(y). \quad (4)$$

We refer to such a functional $U(\cdot)$ as a *no-limit external-control* functional.

We can directly utilize LGH's Theorem 1 to characterize the class of preference relations whose restriction to the set of menus not subject to external control admit a limited self-control functional representation of the form given in expression (3). Below we provide additional properties for a preference relation that entails its restriction to the set of menus subject to external control admitting a no-limit external-control functional representation of the form given in expression (4) with the *same* normative and temptation utility scales that appear in the limited self-control functional representation of its restriction to menus not subject to external control.

⁵ LGH dubbed members of this class (and their associated functionals), *costly self-control with limited willpower preferences (functionals)*. We shall employ the more succinct designation introduced in definition 1.

Axiom 1 [EC-Independence] For any three menus $A(x)$, $B(x')$, $C(x'') \in \mathcal{A}^*$, $A(x) > B(x')$ and $\alpha \in (0, 1)$ implies $\alpha A(x) + (1 - \alpha)C(x'') > \alpha B(x') + (1 - \alpha)C(x'')$.

Axiom 2 [EC-Archimedean] For any three menus $A(x)$, $B(x')$, $C(x'') \in \mathcal{A}^*$, if $A(x) > B(x')$ and $B(x') > C(x'')$ then there exists $\alpha, \beta \in (0, 1)$, such that $\alpha A(x) + (1 - \alpha)C(x'') > B(x')$ and $B(x') > \beta A(x) + (1 - \beta)C(x'')$.

Axiom 3 [EC-Costly-Temptation] For any pair of menus $A(x), B(x) \in \mathcal{A}^*$, $A(x) \succsim B(x)$ implies $(A \cup B)(x) \sim B(x)$.

Axiom 4 [EC-Invariant-Lottery-Preferences] For any pair of lotteries $x, y \in \Delta(Z)$ and any pair of menus A and B such that $\{x, y\} \subseteq A \cap B$, $A(x) \succsim A(y) \implies B(x) \succsim B(y)$.

Axiom 5 [Equivalence-when-Self-Control-Not-Limited] For any lottery $x \in \Delta(Z)$, $\{x\} \sim \{x\}(x)$ and $\{x, y\} \sim \{x, y\}(x)$ for all $y \in S(x)$.

The first two axioms are standard. The reasons for why independence may not hold when the DM's willpower is limited, no longer apply if external control commits her to a particular item in the menu. And standard arguments for why a DM should not have a lexicographic preference for certainty motivate the weak continuity embodied in Axiom 2. Axiom 3 states that if external control commits the DM to selecting a particular item from a menu, then expanding the range of options in that menu by adding items from a better menu also subject to external control committing the DM to select the same item, can not improve the former menu. The intuition is that the larger menu formed by taking the union of the two menus, has in it, the same most tempting alternative from the inferior menu. Axiom 4 stipulates that the ranking between any two lotteries, which are selected through external control, is not influenced by the presence of other lotteries in the menu. Finally, Axiom 5 requires the DM express indifference between any menu not subject to external control and the same menu subject to external control that commits the DM to select the normatively most desired element in that menu, whenever that menu has a single element, or has two-elements for which the DM has sufficient willpower to choose the normatively superior alternative over the more tempting alternative. Notice for such a menu as this, the willpower constraint is not binding, and so the axiom requires the normative evaluation as well as the intensity of temptation cravings of any lottery in the menu are the same whether or not the menu is subject to external-control.

Theorem 1 Fix a binary relation $\succsim$. Suppose it can be represented by a function $U : \mathcal{A} \cup \mathcal{A}^* \rightarrow \mathbb{R}$ whose restriction to $\mathcal{A}$ is a limited self-control functional characterized by the triple $\langle u, v, w \rangle$. Then the following are equivalent.

I The binary relation $\succsim$ satisfies Axioms 1 -5.

II The restriction of U to $\mathcal{A}^*$ is a no-limit external-control functional characterized by the pair $\langle u, v \rangle$.

The full proof appears in appendix below but here we provide a sketch of the main ideas and arguments.

Using the Mixture Space Theorem (Herstein and Milnor 1953), Axioms 1 and 2 imply that the binary relation $\succsim$ restricted on the convex hull of a menu A has a linear representation w_A . Axiom 4 ensures that this representation is menu independent. Furthermore, it follows from the uniqueness up to positive affine transformations of linear functions, we can normalize w , so that $w(\cdot) = u(\cdot) + v(\cdot)$. Thus, using Axiom 5, we establish that U must satisfy expression 4 for all the singleton and two-element menus in $\mathcal{A}^*$. Finally, Axiom 3 allows us to extend this result to all the finite external control menus. Following GP, the extension of the representation from finite external control menus to all compact external control menus is achieved by the continuity of u and v .

Theorem 1 implies that the additional external commitment device that controls the choice made in the subsequent hot state is equivalent to removing the willpower constraint that may have prevented the DM from selecting that item from the menu. From expressions 3 and 4, we know for any $A \in \mathcal{A}$, if $A(x) \succsim A(y)$, for all $y \in A$, then $U(A(x)) \geq U(A)$. This motivates the following definition.

Definition 3 The preference $\succsim$ has a preference for external control at A , if $A(x) \succ A$, for some $x \in A$. The preference $\succsim$ has a preference for external control if $\succsim$ has a preference for external control at some $A \in \mathcal{A}$.

Note that the conclusion in Theorem 1 holds for a preference $\succsim$ whose restriction to $\mathcal{A}$ is represented by GP's temptation preferences. Hence, the following corollary highlights the difference between GP's temptation preferences and LGH's costly self-control with limited willpower preferences.

Corollary 1 Suppose a binary relation $\succsim$ restricted to $\mathcal{A}$ admits a limited self-control representation. Then the preference $\succsim$ has a preference for external control if and only if there exists some $a \in \Delta(Z)$ such that $I(a) \neq \emptyset$.

However, it is important to note that should the preference relation $\succsim$ restricted to $\mathcal{A}$ admit a costly self-control representation of the form introduced and axiomatized by Noor and Takeoka (2010), then the *convexity* of self-control costs they characterize, may entail the existence of some alternative a for which $I(a) \neq \emptyset$ even though $\succsim$ does not exhibit a preference for external control.

4 Beyond external control and the hot-cold empathy gap

One implication we can draw from Theorem 1 is, for any menu A and any alternative $\bar{x}$ that is maximal for that menu, in the sense that $A(\bar{x}) \succsim A(y)$ for all y in A , it must be the case that $A(\bar{x}) \succsim A$. That is, it makes sense to use external control when

the DM is certain about what is the best choice from a given menu A . However, the trade-off between external-control and self-discretion becomes more nuanced if the DM, in the initial (cold state) period, although she knows her preferences over singleton menus, is unsure about the limit of her willpower and the strength of her craving in the subsequent period. In such a situation, the DM faces what (Loewenstein 2000a) refers to as a hot-cold empathy gap.

To illustrate how we might extend our framework to incorporate the possibility of a hot-cold empathy gap, consider a DM who in the initial cold state is *uncertain* what her preferences over $\mathcal{A} \cup \mathcal{A}^*$ might be. In particular, suppose she thinks they could be $\succsim_1$ with probability p and $\succsim_2$ with probability $1-p$. Furthermore, suppose both relations satisfy the hypothesis of Theorem 1 with $\langle u, v_1, w_1 \rangle$ (respectively, $\langle u, v_2, w_2 \rangle$) characterizing a limited self-control functional that represents the restriction of $\succsim_1$ (respectively, $\succsim_2$) to $\mathcal{A}$. If both relations satisfy Axioms 1-5 then from theorem 1 it follows their respective restrictions to $\mathcal{A}^*$ admit no-limit external-control functional representations characterized by $\langle u, v_1 \rangle$ and $\langle u, v_2 \rangle$, respectively.

Finally, suppose $\succsim_1$ exhibits more self-control than $\succsim_2$ as defined in (Liang et al. 2020 Theorem 3), that is, $v_2 = \rho v_1$ and $w_2 \leq \rho w_1$ for some $\rho > 1$. Consider a menu $\{x, y\}$ with $u(x) = v_1(x) = 0, u(y) = -20, v_1(y) = 12, w_1 = w_2 = 10$, and $\rho = 2$. Regardless as to whether nature selects $\succsim_1$ or $\succsim_2$, the DM will give in to temptation and select y in the hot state and have a preference for commitment in the initial state, that is, $\{x\} \succ_i \{x, y\} \sim_i \{y\}$ for $i = 1, 2$.

Notice, however,

$$U_1(\{x, y\}(x)) = u(x) - (v_1(y) - v_1(x)) = -12 \quad \text{and} \quad U_1(\{x, y\}) = u(y) = -20,$$

that is, $\{x, y\}(x) \succ_1 \{x, y\}$, and so we see $\succsim_1$ exhibits a preference for external control at $\{x, y\}$. On the other hand, since

$$U_2(\{x, y\}(x)) = u(x) - (v_2(y) - v_2(x)) = -24 \quad \text{and} \quad U_2(\{x, y\}) = u(y) = -20,$$

we see $\{x, y\} \succ_2 \{x, y\}(x)$, that is, $\succsim_2$ does *not* exhibit a preference for external control at $\{x, y\}$. So should a hard commitment device to implement the menu $\{x\}$ not be available, a choice to commit to an external control device to implement $\{x, y\}(x)$ in the initial cold state *reveals*, p , which recall is her (subjective) assessment that $\succsim_1$ rather than $\succsim_2$ will govern her choice in the subsequent hot state satisfies the inequality

$$p U_1(\{x, y\}(x)) + (1-p) U_2(\{x, y\}(x)) > u(y) \implies p > \frac{1}{3}.$$

More generally, in situations that are not a matter of life or death, and where there is only a relatively small chance that giving in to temptation will be due to insufficient willpower, the DM would value having discretion in the (hot state) subsequent period. Therefore, as the example above illustrates, her choice to commit to an external control device in the initial cold state *reveals* her (subjective) assessment of a relatively high likelihood of experiencing a craving in the hot state for which she will have insufficient willpower to resist.

Although we leave it for future work to extend our model to allow the DM to be uncertain about the strength of her cravings and the limits of willpower, we conclude by noting that such an extension would have distinct positive predictions from the other models in the extant literature. For changing-tastes and conflict between selves modeling, the DM strictly prefers external control to self-discretion regardless as to whether there is a hot-cold empathy gap or not. However, for temptation preferences, like GP and Noor and Takeoka (2010)'s increasing ("up-hill") self-control cost representations, the DM always prefers self discretion to external control. However, as we have argued above, for costly self-control with limited willpower preferences like LGH, which can be seen as the intermediate case of the Strotz and GP representations, the DM in general faces a non-trivial trade-off between external-control and self-discretion.

5 Appendix. Proof of Theorem 1

Step 1. First, fix an arbitrary menu $A \in \mathcal{A}$. We denote the convex hull of A by $\bar{A}$ and define the induced relation $\succeq_A \subset \bar{A} \times \bar{A}$ by setting for each pair of lotteries x and y in $\bar{A}$, $x \succeq_A y$, if $\bar{A}(x) \succeq \bar{A}(y)$.⁶ Notice that $\succeq_A$ inherits the property of completeness and transitivity from the underlying relation $\succeq$. By standard arguments, Axioms 1 and 2 imply the complete and transitive relation $\succeq_A$ admits an expected utility (that is, linear) representation $w_A : \bar{A} \rightarrow \mathbb{R}$. By Axiom 4 it follows for any pair of lotteries $x, y \in \Delta(Z)$ and any pair of menus A and B , $w_A(x) \geq w_A(y)$ if and only if $w_B(x) \geq w_B(y)$. So we can 'piece' together the different cardinal representations for each menu into a single function $w : \Delta(Z) \rightarrow \mathbb{R}$ that 'represents' the DM's preferences over lotteries in the subsequent period when her choice is not constrained by the limit of her willpower.

Step 2. The fact that the restriction of U to $\mathcal{A}$ is a functional of the form given in expression (3) demonstrates that we may identify the utility of any menu without external-control $A \in \mathcal{A}$ with an appropriate two-element subset. The next lemma shows that this extends to any finite $A(x) \in \mathcal{A}^*$.

Lemma 1 Let U be a function that represents a complete and transitive relation $\succeq$ that satisfies Axiom 3 restricted to the set of external control menus $\mathcal{A}^*$. If $A(x) \in \mathcal{A}^*$ is a finite set then $U(A(x)) = \min_{y \in A} U(\{x, y\}(x))$.

Proof Suppose $u^* = \min_{y \in A} U(\{x, y\}(x))$ and $y^* \in A$ is a solution to this problem. By construction $U(\{x, y^*\}(x)) \leq U(\{x, y\}(x))$ for all $y \in A$. Therefore, repeated application of Axiom 3 implies $U(A(x)) = U(\cup_{y \in A} \{x, y\}(x)) = U(\{x, y^*\}(x)) = u^*$. As required. $\square$

Step 3. By Axiom 5, for each lottery $x \in \Delta(Z)$ we can set $U(\{x\}(x)) := u(x)$ for all. Hence, $U(\{x\}(x))$ satisfies expression (4). Next, we will show expression (4)

⁶ It is natural to interpret $x \succeq_A y$ as revealing that the DM prefers to commit to an external control device that selects lottery x rather than lottery y from the menu A .

holds for two-element menus as well. Consider any two-element menu $\{x, y\}$, where x and y are in the relative interior of $\Delta(Z)$ and $v(y) \geq v(x)$. Select a lottery $\bar{x}$ such that $y \in S(\bar{x})$ and $y \in S((1 - \alpha)\bar{x} + \alpha x)$ for sufficiently small α . Hence, by Axiom 5 and Claim 1 below, we have

$$\begin{aligned} U(\{(1 - \alpha)\bar{x} + \alpha x, y\}) &= U(\{(1 - \alpha)\bar{x} + \alpha x, y\}((1 - \alpha)\bar{x} + \alpha x)) \\ &= (1 - \alpha)U(\{\bar{x}, y\}(\bar{x})) + \alpha U(\{x, y\}(x)). \end{aligned}$$

Substituting in for $U(\{(1 - \alpha)\bar{x} + \alpha x, y\})$ and $U(\{\bar{x}, y\}(\bar{x}))$ we get

$$\begin{aligned} (1 - \alpha)(u(\bar{x}) + v(\bar{x})) + \alpha(u(x) + v(x)) - v(y) \\ = (1 - \alpha)(u(\bar{x}) + v(\bar{x}) - v(y)) + \alpha U(\{x, y\}(x)). \end{aligned}$$

Hence,

$$U(\{x, y\}(x)) = u(x) + v(x) - v(y).$$

Note that step 1 establishes that there exists $w : \Delta(Z) \rightarrow \mathbb{R}$ such that for any $\{x, y\} \subset A \subset \mathcal{A}$, $x \succeq_A y$ iff $w(x) \geq w(y)$. Furthermore, it follows from the uniqueness up to positive affine transformations of linear functions, we can normalize w , so that $w(\cdot) = u(\cdot) + v(\cdot)$. Therefore, we conclude $U(\{x, y\}(y)) = u(y) = u(y) + v(y) - v(y)$.

Step 4. The extension of the representation from finite external control menus to all compact external control menus follows from the following lemma (taken from GP, Lemma 0, p1421) and the continuity properties of U .

Lemma 2 If X is a compact set in a metric space, then there exists a sequence of subsets X_k of X such that each X_k has k elements and the sequence X_k converges to X in the Hausdorff metric.

Claim 1 $U(\{(1 - \alpha)\bar{x} + \alpha x, y\}((1 - \alpha)\bar{x} + \alpha x)) = (1 - \alpha)U(\{\bar{x}, y\}(\bar{x})) + \alpha U(\{x, y\}(x))$

Proof Let $A = (1 - \alpha)\{\bar{x}, y\} + \alpha\{x, y\} = \{(1 - \alpha)\bar{x} + \alpha x, (1 - \alpha)\bar{x} + \alpha y, (1 - \alpha)y + \alpha x, y\}$. Hence, by the linearity of U restricted to $\mathcal{A}^*$ (that follows from Axiom 1), we have

$$U(A((1 - \alpha)\bar{x} + \alpha x)) = (1 - \alpha)U(\{\bar{x}, y\}(\bar{x})) + \alpha U(\{x, y\}(x)).$$

By Lemma 1, we have

$$U(A((1 - \alpha)\bar{x} + \alpha x)) = \min_{z \in A} U(\{(1 - \alpha)\bar{x} + \alpha x, z\}((1 - \alpha)\bar{x} + \alpha x)).$$

Suppose $U(A((1 - \alpha)\bar{x} + \alpha x)) = U(\{(1 - \alpha)\bar{x} + \alpha x, (1 - \alpha)\bar{x} + \alpha y\}((1 - \alpha)\bar{x} + \alpha x))$
 $= U((1 - \alpha)\{\bar{x}\}(\bar{x}) + \alpha\{x, y\}(x)) = (1 - \alpha)u(\bar{x}) + \alpha U(\{x, y\}(x)) >$

$(1 - \alpha)U(\{\bar{x}, y\}(\bar{x})) + \alpha U(\{x, y\}(x))$, which yields a contradiction. Furthermore, by $y \in S((1 - \alpha)\bar{x} + \alpha x)$, Axiom 5 and $v(y) > v(x)$, we have

$$\begin{aligned}
& U((1 - \alpha)\bar{x} + \alpha x, y)((1 - \alpha)\bar{x} + \alpha x) \\
&= (1 - \alpha)(u(\bar{x}) + v(\bar{x})) + \alpha(u(x) + v(x)) - v(y) \\
&< ((1 - \alpha)(u(\bar{x}) + v(\bar{x}) - v(y)) + \alpha u(x)) \\
&= (1 - \alpha)U(\{\bar{x}, y\}(\bar{x})) + \alpha U(\{x\}(x)) = U((1 - \alpha)\bar{x} + \alpha x, (1 - \alpha)y + \alpha x)((1 - \alpha)\bar{x} + \alpha x) \\
&< (1 - \alpha)u(\bar{x}) + \alpha u(x) = U((1 - \alpha)\bar{x} + \alpha x).
\end{aligned}$$

□

Funding Open access funding provided by Academia Sinica.

Data availability We do not analyze or generate any datasets, because our work proceeds within a theoretical and mathematical approach. One can obtain the relevant materials from the references below.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

- Ashraf, N. (2009). Spousal control and intra-household decision making: An experimental study in the philippines. *American Economic Review*, 99(4), 1245–1277.
- Bernheim, B. D., & Rangel, A. (2004). Addiction and cue-triggered decision processes. *American economic review*, 94(5), 1558–1590.
- Bernheim, B., & Rangel, A. (2007). Behavioral public economics: welfare and policy analysis with nonstandard decision-makers. In P. Diamond & H. Vartiainen (Eds.), *Behavioral economics and its applications*. Princeton University Press.
- Bryan, G., Karlan, D., & Nelson, S. (2010). Commitment devices. *Annual Review of Economics*, 2(1), 671–698.
- Dekel, E., & Lipman, B. L. (2012). Costly self-control and random self-indulgence. *Econometrica*, 80(3), 1271–1302.
- Fudenberg, D., & Levine, D. K. (2006). A dual-self model of impulse control. *American Economic Review*, 96(5), 1449–1476.
- Gul, F., & Pesendorfer, W. (2001). Temptation and Self-Control. *Econometrica*, 69(6), 1403–1435.
- Gul, F., & Pesendorfer, W. (2004). Self-control and the theory of consumption. *Econometrica*, 72(1), 119–158.
- Herstein, I. N. and Milnor, J. (1953). An axiomatic approach to measurable utility. *Econometrica, Journal of the Econometric Society*, 291–297.
- Laibson, D. (1996). Hyperbolic discount functions, undersaving, and savings policy.
- Liang, M., Grant, S., & Hsieh, S. (2020). Costly self-control and limited willpower. *Economic Theory*, 70, 607–632.
- Loewenstein, G. (2000). Emotions in economic theory and economic behavior. *American Economic Review*, 90(2), 426–432.
- Loewenstein, G. (2000). Willpower: a decision-theorist's perspective. *Law and Philosophy*, 19, 51–76.
- Noor, J., & Takeoka, N. (2010). Uphill self-control. *Theoretical Economics*, 5(2), 127–158.

- O'donoghue, T. and M. Rabin,. (1999). Doing it now or later. *American economic review*,89(1), 103–124.
- Schelling, T. C. (1984). Self-command in practice, in policy, and in a theory of rational choice. *The American Economic Review*, 74(2), 1–11.
- Shefrin, H. M., & Thaler, R. H. (1988). The behavioral life-cycle hypothesis. *Economic Inquiry*, 26(4), 609–643.
- Strotz, R. (1955). Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies*, 23(3), 165–180.
- Thaler, R. H., & Shefrin, H. M. (1981). An economic theory of self-control. *Journal of Political Economy*, 89(2), 392–406.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.