Temptation and Guilt*

Jawwad Noor *Linxia Ren*

June 26, 2022

Abstract

An agent experiences guilt when she submits to temptation, and the anticipation of this can generate guilt-avoidance, such as crossing the street to avoid being solicited for a charitable donation. This paper builds a model of guilt on the observation that guilt-avoidance may itself cause guilt. The model generates this by treating the guilt-avoidance motive as a temptation. Behavioral foundations and welfare implications are explored. Applications to social preferences, perfectionism and information avoidance are presented.

Keywords: Guilt, normative preference, temptation, social preference, perfectionism, welfare.

1 Introduction

Temptation arises when desires (*temptation preference*) conflict with how one feels they should behave (*normative preference).* A key implication of

*Noor (email: jnoor@bu.edu) is at the Dept of Economics, Boston University, 270 Bay State Road, Boston MA 02215 and Ren (email: renlinxia.job@gmail.com) is at TikTok, 250 Bryant St, Mountain View, CA 94041. We thank the associate editor, two referees, Bart Lipman, Philipp Sadowski, Mike Luca, Rob Munger, Sean Horan, seminar participants at Boston University, Columbia University and University of Rochester, and participants at the Far East Econometric Society Meeting (2009) for useful comments. Ren gratefully acknowledges financial assistance from the Institute of Economic Development at BU. The usual disclaimer applies.

self-control problems identified in the large literature on the topic (Strotz [42], Laibson [33], Gul and Pesendorfer [26, 27], Fudenberg and Levine [23, 24, 25], Noor and Takeoka [38]) is that such internal conflict will give rise to a *preference for commitment*. For instance, if the agent believes that she should work (denoted w) but is tempted to respond to her text messages (denoted m), then she might turn off notifications on her cellphone to avoid facing temptation while working, thereby exhibiting the following preference over menus (choice problems):

$$\{w\} \succ \{w, m\}.$$

In the literature, a preference for commitment is used to uncover an agent's (subjective) normative and temptation preferences (Gul and Pesendorfer [26, 27]). The application to welfare policy is immediate: we can learn what might improve people's welfare by observing their commitment preferences. If people use internet blockers to help them work, we learn that more effective internet blockers should only make them better off. Della Vigna et al [14] show that subjects in their field experiment use a "do not disturb" option (which serves as a commitment option) to avoid being contacted in person for donations, the authors conclude that door-to-door solicitations for donations are in fact welfare-reducing for the donors.

This paper models the psychological phenomenon of *guilt*. Overturning received wisdom, our agent may in fact commit to "bad" alternatives. Imagine that the agent in the opening example is in fact a workaholic – she is tempted to work in the evening while also considering it normatively better to interact with her friends and family who are available at that time. Anticipating that she will give into the temptation to work w, she may recognize that she would feel guilty if she does not respond to text messages m while she works. *Guilt-avoidance* may then induce the preference for commitment given above, so that she can indulge her temptation to work guilt-free.[1] Similarly, people may genuinely adhere to the normative standard of being charitable, but are in fact tempted to be selfish. By using a "do not disturb" option, they can indulge their selfishness without being reminded of how they may be failing their normative standards. Contrary to the conclusion of Della Vigna et al [14] and the literature more generally, in our story the "do not disturb" option may be leading agents to make

[1]Noor [36, 37] shows that an agent may fail to commit despite the presence of temptation (due to temptation by menus). This paper makes a bolder claim that a temptation-stricken agent may in fact commit to indulgence.

choices that are *welfare-reducing*. Indeed, door-to-door solicitations might be welfare-enhancing because social-pressure *helps* them behave in accordance with their normative standards.[2] This point of view also suggests a reinterpretation of the literature on moral hypocrisy (Batson et al [5]) that takes the position that true preference is selfish, and non-selfish behavior arises only for appearances.

A interesting implication of guilt is that *the avoidance of guilt is itself accompanied with guilt*.[3] Thus, although the agent may relieve her guilt ex post when facing her menu, she may experience guilt ex ante at the time that she turns off notifications on her cellphone. Similarly, she may experience guilt when exercising the "do not disturb" option.[4] This gives rise to an obvious infinite regress, implying that temptation and guilt impinge on behavior in any period that we observe the agent. This raises the question of how to tease apart normative and temptation preferences from behavior. We address this by appealing to the idea prevalent in the literature that immediate temptation is stronger than more distant temptation. With "sufficient distance" (which can be thought of in terms of the limit of a sequence of preferences as in Noor [37]), behavior should be close to normative: the workaholic may turn off notifications for the evening but would make future plans with friends and family that force her to forgo working in the evening. Thus, normative preference can be identified by using "sufficient distance" as a tool, and the gap between normative preference and observed behavior can be attributed to temptation and guilt. While modelling guilt raises a unique set of interesting issues we therefore see that, perhaps unexpectedly, existing tools in the literature on temptation are enough to deliver a coherent model of guilt.

In terms of behavioral implications we find that our model gives rise to the following:

1. The model explains why people may exhibit virtuousness (publically

[2]It is not the aim of this paper to resolve whether the "do not disturb" option is an instance of a preference for commitment or that of a commitment to indulgence. This is left to future research.

[3]This feature is not shared by most other painful experiences. For instance, if risk is painful to experience, then the avoidance of risk does not induce any pain.

[4]A host of such examples can be constructed. For instance, if a person feels he should confess a transgression to a friend, then facing his friend and not confessing will cause guilt, but so will purposefully avoiding him (we thank a referee for suggesting this example). Similarly, refusing a solicitation by a representative from a charity may cause guilt, but so may crossing the road in order to avoid one.

announcing a plan to quit smoking, for instance) or make ambitious plans (obtaining an annual membership at a gym) even though they do not subsequently follow through with this. We show that these behaviors can be viewed as strategies to improve subsequent choices: guilt dampens the attractiveness of temptation and helps keep them moderate. We also show that in the interim period the agent would avail opportunities to undo past choices if possible and settle for mediocre options in order to avoid the guilt due to the ambitious plans/standards. Our model provides a unifying explanation for behaviors associated with perfectionism (Kopylov [30]).

2. We show that guilt avoidance can give rise to an ex ante preference for ignorance when information exacerbates temptation. Similarly a preference for information exists when ignorance exacerbates temptation.

3. Experiments based on the dictator game have produced a host of findings on social preferences. As noted above, research suggests that people behave morally because they desire to *appear* moral – indeed, studies find that if agents can avoid having to behave morally and pursue self-interest instead, they do. This view of people as moral hypocrites has gained favor in the social preferences literature (for instance see Neilson [35], Andreoni and Bernheim [3], Della Vigna et al [14]). We show that our model adapted to the social setting (by positing a temptation to be selfish) naturally accommodates the key comparative statics observed in experiments.

While various models in the literature explain each of the above findings individually, we show that these can be unified by our model. But in particular, as noted above, our model has starkly different welfare implications than models in the literature.

The remainder of the paper is organized as follows. Section 2 provides an introspective discussion of guilt around which an axiomatic model is built in Section 3. Section 4 provides a behavioral definition of guilt-proneness. Sections 5-8 are devoted to applications and Section 9 discusses related decision-theoretic literature. Section 10 concludes. All proofs are contained in appendices.

2 Modelling Guilt

2.1 Conceptual Foundations

In this section we describe the intuitions that guide our subsequent modelling choices and axioms.

Let d stand for 'donate' and $\neg d$ for 'do not donate'. Consider an agent who sees a representative from a charitable organization up ahead, because of which he will face the menu $\{d, \neg d\}$. He anticipates experiencing guilt due to the anticipated choice of $\neg d$ from this menu. He seeks to avoid this preemptively by crossing the road, where there is no one soliciting donations, and thus where his menu becomes $\{\neg d\}$. Since there is no opportunity to donate there is no anticipated guilt. We argue, however, that we should expect the agent to feel guilty for crossing the road. Choosing $\{\neg d\}$ violates the same normative desire to donate that causes the anticipated guilt from choosing $\neg d$ in $\{d, \neg d\}$. Thus, *guilt-avoidance should be accompanied with guilt*. This example is illustrated in the following preferences, where time 1 refers to the point where the agent chooses a menu (whether to cross the road or not), and time 2 refers to the time of choice from a menu.

$$\{\ \overset{\text{no guilt at } t{=}2}{\downarrow}\ \neg d\ \underset{\text{guilt at } t{=}1}{\uparrow}\ \} \succ \{d,\ \overset{\text{guilt at } t{=}2}{\downarrow}\ \neg d\ \}.$$

Note that guilt-avoidance here refers to a choice between menus, and thus there is a distinction to be made between the guilt experienced when choosing between menus (time 1) versus from a menu (time 2).

From the preceding observation, we infer the following.

Observation N. *Guilt-avoidance is not a feature of time 1 normative preference.*

Just as the agents normative perspective determines whether to donate or not at time 2, it would determine what menu is better or worse at time 1. If the agent believed that it is normatively justified to cross the road in order to avoid guilt, then he would not experience guilt at time 1. Thus, guilt due to choosing $\{\neg d\}$ over $\{d, \neg d\}$ implies that guilt-avoidance is not a normative goal.

To push the point further we argue that guilt-avoidance would have counter-intuitive implications if it were a normative goal. If an agent normatively preferred committing to a bad action in order to avoid guilt, we

might expect him to also normatively prefer committing to a slightly worse action. That is, $\{bad\} \succ \{good, bad\}$ would imply $\{bad - \varepsilon\} \succ \{good, bad\}$ for some small ε, in an appropriate sense. But it is neither intuitive nor compelling that committing to choose a normatively worse alternative could be deemed normatively desirable when every alternative in the binary menu is normatively better.

Observation T. *Guilt-avoidance is a time 1 temptation.*

Instead of being a normative motivation for crossing the road, we posit that guilt-avoidance may in fact be motivated by temptation. Indeed, introspection confirms that guiltless indulgence is more tempting than guilt-ridden indulgence. For instance, violating ones diet may be less enjoyable in the presence of a health-conscious friend. We conclude that menus that provide greater indulgence with less guilt may be more tempting. Note that just as alternatives may tempt within a menu, menus themselves may tempt. While *bad* may be a more tempting alternative than *good*, the menu $\{bad\}$ would be more tempting than the menu $\{good, bad\}$ which in turn would be more tempting than the menu $\{good\}$.

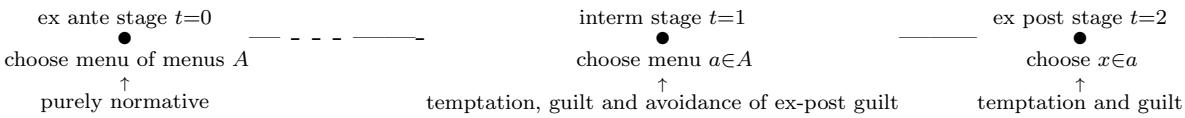
These comments form the conceptual basis for the model we construct.

2.2 Domain

The preceding analysis suggests that in order to capture some of the behavioral impacts of time 2 guilt we may need to account for a temptation by guilt-avoidance, and in particular we need a model that permits *tempting menus* (Noor [37, 36]). Just as Gul and Pesendorfer (2001) use a time 1 preference over menus to capture time 2 temptation, a three period version of their model would allow us to use a time 0 preference over menus of menus in order to capture time 1 temptation by menus (and also time 2 temptation).

To be more explicit, there are three periods: ex ante, interim and ex post. In the ex post period the agent makes a consumption choice *from* some menu. In the interim stage the agent makes her choice *of* menu. In the ex ante stage the agent chooses the set of menus she will face in the interim period. For instance, imagine choosing a neighborhood, then selecting a restaurant in it, and then finally choosing a dinner item at the restaurant. Or consider the fact that by obtaining a gym membership, the set of possibilities available after work expands to include going to the gym (going home, or to the gym, or to a restaurant is each a menu of options), and one subsequently selects an option from this set and subsequently makes their final consumption choice.

A key assumption we make is that *choice in the ex ante stage is prior to the experience of temptation, and choice in the remaining stages is subject to temptation.* That is, the agent is in a "cold" state in the ex ante period and "hot" state in subsequent periods.



Due to this assumption, the ex ante preference reveals the agent's normative and temptation preferences over menus (and in turn over final outcomes). For instance, if the agent prefers to commit to an interim menu a rather than b,

$$\{a\} \succsim \{b\},$$

then this suggests that from her ex ante (normative) perspective a is better. Thus such "commitment preferences" identify the agent's normative preference over menus. If the agent exibits an ex ante preference for commitment,

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

then it is revealed that a is normatively superior to b and that b is more tempting than a. In the next section we write axioms that aim to capture how temptation by guilt-avoidance is reflected in a preference over menus of menus.

One might object to the assumption that the agent can be in a cold state in the ex ante period. Consider the illustration in the previous section. Anticipating that at time 2 she will submit to the temptation to choose $\neg d$ and feel guilty about it, in time 1 the desire to avoid guilt leads her to cross the street in order to avoid a representative from a charity, captured by the choice of menu $\{\neg d\}$ over menu $\{d, \neg d\}$. Our model will accordingly visualize the time 1 choice of $\{\neg d\}$ as the result of a temptation to avoid guilt, and thus this choice will give rise to guilt at time 1. But how is this to be captured? We have suggested to add a time 0 choice of menu of menus. But one might expect temptation and guilt to arise in period 0 as well: for instance, she may choose a neighborhood where there are no representatives on any street $\{\{\{\neg d\}\}\}$ rather than one where there are representatives on some streets $\{\{d\}, \{d, \neg d\}\}$, and this may be a temptation and a cause for guilt. This in fact should lead to an infinite regress, with the implication that the agent must be in a hot state in every period, that is, it suggests

that we can never observe an agent's choices in a cold state, rendering our primitive non-observable.

This is related to the problem addressed in Noor [37], and our solution is the same: we argue that the choices an agent would make in a cold state can be *inferred* from temptation-stricken choices *if* we hypothesize that temporal distance weakens temptation. This is introspectively plausible: the temptation to indulge a craving may be stronger when one is already in a restaurant and close to the time of consumption, compared to at a prior time when one is choosing the restaurant. Indeed, the cushion afforded by the temporal distance may enable one to go to a salad chain and avoid facing temptation altogether. We argue that such considerations hold for guilt as well: the temptation to cross the street – and the guilt from crossing the street – may be higher when the charity representative is close, as opposed to an earlier time when the representative is a more distant thought. As in Noor [37], this allows us to argue that the ranking of "infinitely distant" menus (a ranking that is derived as the limit of a sequence of preference of $t-$period delayed menus) describe the agent's cold-state preference. In this paper we circumvent the technicalities of an infinite horizon model with temptation and guilt in every period. We do so by imagining that period 1 is "infinitely distant" from period 0, in the sense that the period 0 ranking of period 1 menus is not subject to temptation and thus causes no guilt. The "full blown" infinite horizon version of our model corresponding to Noor [37] is presented in Appendix A.

3 Foundations

For any compact metric space Z, let $\Delta(Z)$ denote the set of probability measures on the Borel σ-algebra of Z, endowed with the weak convergence topology; $\Delta(Z)$ is compact and metrizable [2], and we often write it simply as Δ with generic elements x, y, z, and we often refer to these are *alternatives*. Let $\mathcal{M}_1 = \mathcal{K}(\Delta)$ denote the set of all nonempty compact subsets of Δ, with generic elements a, b, c. When endowed with the Hausdorff topology, $\mathcal{M}_1$ is a compact metric space. An element $a \in \mathcal{M}_1$ is referred to as an *interim menu*. Let $\mathcal{M}_0 = \mathcal{K}(\mathcal{M}_1)$ denote the set of all nonempty compact subsets of $\mathcal{M}_1$, endowed with the Hausdorff topology. An element $A \in \mathcal{M}_0$ is referred to as an *ex ante menu*. Both $\mathcal{M}_0$ and $\mathcal{M}_1$ are compact (Aliprantis and Border [2,

Theorem 3.71]) and the mixture operations in these spaces are continuous.[5]

As determined earlier, our primitive is a preference relation $\succsim$ on $\mathcal{M}_0$.[6] The interpretation is that the agent chooses a menu of menus $A \in \mathcal{M}_0$ in the ex ante stage, subsequently selects a menu $a \in A$ in the interim stage, and then picks final consumption $x \in a$ in the ex post stage. See the timeline presented earlier. Choice in the ex ante stage is prior to the experience of temptation, and choice in the remaining stages is subject to temptation. Guilt that may be experienced in the ex post stage affects what menus tempt in the interim stage.

Suitably adapted versions of GP's axioms imposed on $\succsim$ characterize the following basic representation theorem (see Kopylov and Noor [31]). In the interest of brevity, we simply assert the existence of this representation as an axiom. Say that a function $f : X \to \mathbb{R}$ is *linear* if for all $\alpha \in [0,1]$ and $x, y \in X$,

$$f(\alpha x + (1-\alpha)y) = \alpha f(x) + (1-\alpha) f(y),$$

and similarly for any function on $\mathcal{M}_1$.

Axiom 1 (Basic) *The preference $\succsim$ has a utility representation $W : \mathcal{M}_0 \to \mathbb{R}$ such that for all $A \in \mathcal{M}_0$ and $a \in \mathcal{M}_1$,*

$$W(A) = \max_{a \in A} [U(a) - \left[\max_{b \in A} V(b) - V(a) \right]] \tag{1}$$

$$s.t.\ U(a) = \max_{x \in a} [u(x) - \left[\max_{y \in a} v(y) - v(x) \right]], \tag{2}$$

where $u, v : X \to \mathbb{R}$ and $V : \mathcal{M}_1 \to \mathbb{R}$ are continuous linear functions.

The ex ante preference $\succsim$ over menus of menus admits the representation (1), which is defined in terms of the agent's normative and temptation preferences over *menus*, represented by U and V respectively. The desired functional form (2) for normative utility U is asserted and interpreted shortly.

[5]For $\alpha \in [0,1]$, let $\alpha x + (1-\alpha)y \in \Delta$ denote the α-mixture that assigns $\alpha x(S) + (1-\alpha)y(S)$ to each S in the Borel $\sigma-$algebra of Z. Similarly, let $\alpha a + (1-\alpha)b \equiv \{\alpha x + (1-\alpha)y : x \in a, y \in b\} \in \mathcal{M}_1$ denote an α-mixture of menus a and b. Finally, an α-mixture of menus of menus A and B is given by $\alpha A + (1-\alpha)B \equiv \{\alpha a + (1-\alpha)b : a \in A, b \in B\} \in \mathcal{M}_0$.

[6]This choice domain is used in Kopylov and Noor [31]. An important difference between this domain and the domain of multi-period menus used in GP [27] is the absence of lotteries over menus. This feature is obtained by exploiting Kopylov [29].

However, the desired functional form for V has yet to be ascertained, and will be derived from axioms in the sequel.

To interpret normative utility (2) over final outcomes: when evaluating an interim menu a the agent considers the value of normative ex post utility $u(x)$ net of self-control costs $[\max_{y\in a} v(y) - v(x)]$ attained by ex post choice from that menu. Thus, the agent cares only about the normative utility of ex post choice and the self-control costs of obtaining it. For later reference we note that the functional form suggests that anticipated ex post choice $c(a)$ is given by

$$c(a) = \arg\max_{x\in a}\left(u(x) - \left[\max_{y\in a} v(y) - v(x)\right]\right) = \arg\max_{x\in a}\left(u(x) + v(x)\right).$$

Note that a guilt term such as $[\max_{y\in a} u(y) - u(x)]$ does not feature in (2). If it did we would have a model where guilt hurts the normative value of a menu and thus leads to a normative preference to avoid guilt. But we determined in Observation N (Section 2) that there cannot be a normative motivation to avoid guilt. *This is the reason that GP's model (2) can be applied for normative preference without any need for modification.*

Next, the ex ante (normative) utility of a menu of menus given in (1) has exactly the same form as in (2), where the agent cares about the interim choice and the normative utility $U(a)$ from the interim menu net of self-control costs $[\max_{b\in A} V(b) - V(a)]$. As before, interim choice is the argmax of $U+V$ and interim guilt $[\max_{b\in A} U(b) - U(a)]$ does not feature here either, again because of Observation N.

Ex post guilt, however, is relevant for temptation preference V over menus, given that we visualize this temptation as being motivated to avoid guilt as in Observation T. We proceed to write axioms that reflect this and then obtain a functional representation for V as the result of a theorem.

3.1 Main Axioms

We augment the basic model with three behavioral assumptions. The following interpretation of behavior are needed to interpret the assumptions. The reader should keep track of whether each is a statement about temptation or choice in the interim or ex post period.

- $\{a\} \succ \{a, b\}$ reveals that in the interim stage, menu b tempts menu a.
- $\{a\} \succ \{a \cup b\}$ reveals that in the ex post stage, an alternative in b is a source of temptation in menu $a \cup b$. Observe that interim choice is trivial

in both $\{a\}$ and $\{a \cup b\}$. Therefore the preference reveals information about what is experienced in the ex post period.

- $\{a \cup b\} \succ \{b\}$ reveals that ex post choice from $a \cup b$ belongs to a. Observe that if the anticipated choice from $a \cup b$ lay in b, then there would be no reason to strictly prefer committing to the larger menu $a \cup b$.

Our first behavioral assumption is a consistency requirement.

Axiom 2 (Temptation Consistency) *For* ***singleton*** *interim menus* $a, b \in \mathcal{M}_1$,

$$\{a\} \succ \{a, b\} \iff \{a\} \succ \{a \cup b\}.$$

This is an innocuous consistency requirement that ensures that there is no wedge between what final consumption the agent finds tempting in the interim period and in the ex post period. In the context of the donate d and not-donate $\neg d$ example, the axiom states $\neg d$ is an interim temptation ($\{\{d\}\} \succ \{\{d\}, \{\neg d\}\}$) if and only if it is an ex-post temptation ($\{\{d\}\} \succ \{\{d, \neg d\}\}$).

The next behavioral assumption expresses the idea that temptation preference over menus is 'forward-looking' in the sense of being sensitive to ex post choice. Observe that sensitivity to ex post choice is necessary for menu-temptation to be sensitive to ex post guilt, as in Observation T (Section 2).

Axiom 3 (Temptation Sophistication) *For any interim menus* $a, b \in \mathcal{M}_1$,

$$\{a \cup b\} \succ \{b\} \Longrightarrow \{a\} \not\succ \{a, a \cup b\}$$

The axiom states that if the agent anticipates the ex post choice from $a \cup b$ to lie in a, then in the interim stage the agent can never be tempted by the menu $a \cup b$ when a is available. In particular, this is true even if b contains something tempting. This is expressed rather directly in the contrapositive: if the agent is tempted to keep the option $\neg d$ in the interim period (revealed by an ex ante preference for committing to donating, $\{\{d\}\} \succ \{\{d\}, \{d, \neg d\}\}$), then it must be that she would choose $\neg d$ in the ex post period if she could ($\{\{d\}\} \succ \{\{d, \neg d\}\} \sim \{\{\neg d\}\}$).[7] The axiom captures the idea that a menu

[7]Note that by Basic, $\{\{d\}\} \succ \{\{d\}, \{d, \neg d\}\}$ implies $\{\{d\}\} \succ \{\{d, \neg d\}\}$, and the contrapositive of Temptation Sophistication requires $\{\{d, \neg d\}\} \precsim \{\{\neg d\}\}$ and so we obtain $\{\{d\}\} \succ \{\{d, \neg d\}\} \sim \{\{\neg d\}\}$.

tempts in the interim period by virtue of what is chosen in the menu ex post, and not its tempting content per se.

The next axiom is the substantive one that relates temptation by menus to guilt in a manner consistent with Observation T. In particular, it behaviorally expresses temptation to avoid guilt. Define the relation $\succ_0$ over $\mathcal{M}_1$ by

$$a \succ_0 b \text{ if there exists } x \in a \text{ s.t. } \{\{x\}\} \succ \{\{y\}\} \text{ for all } y \in b.$$

That is, $a \succ_0 b$ if the 'most virtuous' alternative in a is strictly better than that in b according to the ex ante normative perspective.

Axiom 4 (Guilt-Averse Temptation) *For all interim menus $a, b \in \mathcal{M}_1$ such that $b \subset a$,*

$$\{a\} \succ \{a, b\} \Longrightarrow a \succ_0 b.$$

The axiom states that the temptation to choose a commitment option $b \subset a$ arises only if it excludes virtuous alternatives in a. That is, if there is a temptation to avoid a menu, it must in fact be a temptation to avoid facing a virtuous alternative, and thus guilt. For instance, if the agent is tempted to cross the road ($\{\{d, \neg d\}\} \succ \{\{d, \neg d\}, \{\neg d\}\}$) then it must be because donating is normatively superior to not-donating.

A final axiom is used to characterize the special case of our model that embodies the idea that temptation weakens with distance.

Axiom 5 (Preference for Early Choice) *For all **singleton** interim menus $a, b \in \mathcal{M}_1$,*

$$\{a\} \succ \{a \cup b\} \succ \{b\} \implies \{a, b\} \succ \{a \cup b\}.$$

Consider $\{\{d, \neg d\}\}$ where the agent has no interim choice and must choose between d and $\neg d$ ex-post. Consider also $\{\{d\}, \{\neg d\}\}$ where the agent effectively chooses between d and $\neg d$ in the interim period. Both scenarios involve the choice of final consumption, but one involves a choice in the final period and the other in the interim period. The axiom says that if the agent anticipates resisting temptation ex post and choosing d over $\neg d$, then she would rather face $\{\{d\}, \{\neg d\}\}$ than $\{\{d, \neg d\}\}$. Intuitively, it is harder to resist temptation in the ex post period than it is in the interim period.

3.2 Representation Result

Say that $\succsim$ is *nondegenerate* if there are $A, B \in \mathcal{M}_0$ and $a, b \in \mathcal{M}_1$ such that $A \succ A \cup B \succ B$ and $\{a\} \succ \{a \cup b\} \succ \{b\}$. That is, the agent anticipates resisting temptation in some $a \cup b \in \mathcal{M}_1$ and $A \cup B \in \mathcal{M}_0$.

Theorem 1 *A nondegenerate preference $\succsim$ satisfies Basic, Temptation Sophistication, Temptation Consistency and Guilt-Averse Temptation if and only if $\succsim$ has a utility representation (1)–(2) such that for all $a \in \mathcal{M}_1$,*

$$V(a) = \kappa \max_{x \in a} [v(x) - \left[\max_{y \in a} u(y) - u(x)\right]], \tag{3}$$

where $\kappa > 0$ and u, v are affinely independent. Moreover, $\kappa < 1$ iff Preference for Early Choice holds.

The proof is presented in the appendix.[8] The theorem states that under the three axioms in addition to Basic, we obtain a functional form for V that suggests that the temptation utility of a menu is determined by the temptation utility $v(x)$ of the ex post choice and the guilt cost $[\max_{y \in a} u(y) - u(x)]$ associated with it. This is the GP form with the roles of u and v reversed. The parameter κ is the temptation discount factor applied in period 1 when evaluating utility that is to be received in period 2. A normative discount factor does not appear in Basic since it normalized to 1. When $\kappa < 1$, the model has the property that, with delay, temptation utility is discounted at a steeper rate than normative utility.

It is worth noting that the anticipated choice according to V is exactly the same as that according to U:

$$\arg\max_{x \in a} [v(x) - \left[\max_{y \in a} u(y) - u(x)\right]] = \arg\max_{x \in a} [u(x) + v(x)] = \arg\max_{x \in a} [u(x) - \left[\max_{y \in a} v(y) - v(x)\right]].$$

[8] An outline is as follows. The key step is to prove that for any interim menus a, b,

$$\max_a u = \max_b u \text{ and } \max_a (u+v) = \max_b (u+v) \text{ and } \max_a v = \max_b v \implies V(a) = V(b).$$

Thus, for instance, if a, b have the same normative-best alternative, the same ex post choice and the same ex post temptation, then the menus are equally tempting. This is proved by contradiction: if $V(a) \neq V(b)$ then we can find menus a^*, b^* that are not consistent with Temptation Consistency and Guilt-Averse Temptation. By Harsanyi's aggregation theorem, we obtain that $V(a) = \kappa \max_a (u+v) + k_2 \max_a v + k_3 \max_a u$ for some scalars κ, k_2, k_3. The remainder of the proof shows that these scalars have appropriate signs that permit us to write V in the desired form.

This expresses the Sophistication axiom, and presumes that temptation preferences are sensitive to anticipated future choice, and in particular are not naive as is assumed in other papers (such as Noor [37, 36]).

In Appendix A we formulate the model in the manner of Noor [37], which explicitly exposits the agent's temptation and guilt experience in every (non-zero) period. In our three period model, the temptation and guilt experiences in periods 1 and 2 are not explicit because we have taken the period 0 preference to reflect a purely normative perspective, and we infer from it what is anticipated about periods 1 and 2. Indeed, the model captures period 2 guilt costs and period 1 temptation to avoid guilt (observation T) only through the functional form for V. There is no guilt term in the model that reflects period 1 guilt costs as a result of guilt-avoiding choices, since the period 0 representation reflects a normative perspective, and by observation N, this is indifferent to any period 1 guilt. Period 2 guilt shows up only because it matters for V, and the period 0 representation takes this into account.

We close by confirming that our representation has the desired uniqueness properties.

Theorem 2 *If $\succsim$ is nondegenerate preference that admits a representation (1)–(3), then it has another representation with components $\kappa' > 0$ and $u', v' \in \mathcal{U}$ if and only if $\kappa' = \kappa$, $u' = \alpha u + \beta_u$ and $v' = \alpha v + \beta_v$ for some $\alpha > 0$ and $\beta_u, \beta_v \in \mathbb{R}$.*

4 Guilt-Proneness

In this section we seek to understand what guilt-proneness means in our formal model. We proceed by defining a comparative behavioral notion of guilt-proneness and then characterizing it. We find that guilt-proneness is equivalent to the tendency to exert self-control.

Consider two agents, $\succsim$ and $\succsim^*$, both of whom satisfy our axioms, and who are *ex post similar* in that they have identical normative and temptation preferences over final consumption: for all $x, y \in \Delta(Z)$,

$$\{\{x\}\} \succsim \{\{y\}\} \iff \{\{x\}\} \succsim^* \{\{y\}\}, \text{ and } \{\{x\}\} \succ \{\{x,y\}\} \iff \{\{x\}\} \succ^* \{\{x,y\}\}.$$

Since our model identifies guilt through temptation by menus, guilt-proneness is naturally identified with the lower tendency to be tempted by menus in which guilt is experienced. To formalize this idea, say that

Definition 1 ([Guilt-Proneness) $\succsim$ *is more guilt-prone than* $\succsim^*$ *if for any menu* $b \in \mathcal{M}_1$ *and* ***singleton*** *menu* $a \in \mathcal{M}_1$,

$$\{a\} \succ \{a, a \cup b\} \Longrightarrow \{a\} \succ^* \{a, a \cup b\}.$$

By Set Betweenness, $\{a\} \succ \{a, a \cup b\}$ implies $\{a\} \succ \{a \cup b\}$. That is, the most tempting alternative in b is more tempting than the alternative in a (recall that a is a singleton). Note that adding the alternative in a to the menu b will then not change the maximum temptation, but it may improve on the normatively-best alternative. Suppose that the agent exhibits $\{a\} \succ \{a, a \cup b\}$, which says that $a \cup b$ tempts a. The representation implies that the chosen alternative in $a \cup b$ cannot be the alternative in a.[9] That is, adding the alternative in a to menu b does not change the choice in the latter. However, the improvement in the normative content can then only *reduce* the temptation value of $a \cup b$ relative to b, as it leads the agent to bear a potentially greater degree of guilt. The definition says that when $a \cup b$ tempts the agent $\succsim$ despite this, then $a \cup b$ will also tempt the less guilt-prone agent $\succsim^*$. The contrapositive states that if $a \cup b$ 'normatively improves' on b to the extent that an agent $\succsim^*$ ceases to be tempted by it, then the more guilt-prone agent $\succsim$ will also cease to be tempted by it. The next theorem characterizes this definition.

Theorem 3 *Let* $\succsim$ *and* $\succsim^*$ *be an ex post similar pair of nondegenerate preferences with guilt representations* (u, v, κ) *and* (u^*, v^*, κ^*) *respectively. Then the following statements are equivalent.*

(a) $\succsim$ *is more guilt-prone than* $\succsim^*$ *.*

(b) Without loss of generality, $v = v^*$ *and* $u = \lambda u^*$ *and* $\lambda \geq 1$*.*

Thus, in our model, to be guilt-prone is to place a higher weight on normative preferences in ex post decisions. That is, guilt-proneness is equivalent to a higher 'intensity' of normative preference. This is intuitive: an agent who experiences guilt more strongly will place more importance on her normative preference, and conversely, an agent who puts high value on her normative preferences is also more prone to feel pangs of guilt if she deviates from them.

Observe that agents who place higher weight on normative preferences will also tend to exhibit a higher degree of self-control in ex post choices –

[9]Suppose it is. Then the choice in a and $a \cup b$ are the same, but the choice in a is guiltless because the menu is a singleton, whereas the choice in $a \cup b$ is potentially guilt-ridden. Therefore, $a \cup b$ cannot possibly tempt a if the choice is the same in both menus.

recall that the agent maximizes $u + v$, and so if u is 'relatively more intense than v' then choices will follow it more closely. Therefore, we see that *greater guilt-proneness is equivalent to exhibiting greater ex post self-control.* This is confirmed in the next theorem. Note that the behavioral definition of 'more ex post self-control' is familiar from GP, and states that whenever $\succsim^*$ is able to resist temptation, then so is $\succsim$.

Theorem 4 *Let $\succsim$ and $\succsim^*$ be an ex post similar pair of nondegenerate preferences with guilt representations (u, v, κ) and (u^*, v^*, κ^*) respectively. Then the following statements are equivalent.*

(a) $\succsim$ is more guilt-prone than $\succsim^*$.

(b) $\succsim$ has more ex post self-control than $\succsim^*$, that is, for all $x, y \in \Delta(Z)$,

$$\{\{x\}\} \succ^* \{\{x, y\}\} \succ^* \{\{y\}\} \Longrightarrow \{\{x\}\} \succ \{\{x, y\}\} \succ \{\{y\}\}.$$

A curious feature of our model is its lack of parameters – there is only one, namely, κ. Our analysis reveals that in our model the notion of guilt-proneness is intimately tied with the notion of self-control, and in particular not controlled by any separate parameter. Intuition strongly suggests that it should not be any other way: it does not seem meaningful to expect to be able to change an agent's sensitivity to guilt without also affecting her tendency to resist temptation. While the previous section explored the testable implications of sensitivity to guilt on an agent's ranking of menus of menus, these results highlight a testable implication for ex post choice, namely, the existence of self-control.

It is interesting to note that higher guilt-proneness leads to both less guilt in period 2 (due to more virtuous choices) but a stronger period 1 temptation to avoid any existing guilt. That is, the guilt-prone agent is often observed to behave virtuously, but in the few situations where she does not, she seems to exhibit nonvirtuous tendencies to commit to indulgence.

We conclude by noting that the parameter κ can be interpreted in terms of how, in the interim period, the agent discounts the temptation of a menu relative to its normative value. Smaller values of this parameter mean that temptation by menus gets less weight in interim decisions, which therefore is expressed behaviorally in the form of greater self-control in the interim stage.

5 Application: Exhibiting Virtuousness

People often make heroic plans that they do not always follow through with. For instance, they may obtain a gym membership which they do not make adequate use of – DellaVigna and Malmendier [15] show that health-club members in their data set could save over $300/year by switching from an annual or monthly contract to per-visit passes. More anecdotally, they make plans to quit smoking, curb spending, etc that are rarely kept. Such behaviors have been explained in the literature in terms of naivete about self-control problems (O'Donoghue and Rabin [39]), and also in terms of an intrinsic preference for adding virtuous alternatives feasible (Kopylov [30]). We show that virtuous plans may have *instrumental value* in dealing with self-control issues, even if agents fail to follow through with them.

To illustrate, suppose that an agent would ordinarily spend his evenings at home and eat regular food $\neg h$, thereby facing the menu $\{\neg h\}$. Alternatively, he could expand his menu to $\{h, \neg h\}$ by keeping healthy alternatives h at home as well. Thus, he faces a menu of menus $\{\{\neg h\}, \{h, \neg h\}\}$. Additionally, he may enroll in a gym, giving him an additional healthier option e of eating regular *and* exercising. (To keep the story simple, we assume away other natural options, such as healthy eating and exercise, or of keeping only healthy food at home). Joining the gym transforms his menu of menus into $\{\{\neg h, e\}, \{h, \neg h, e\}\}$.

Proposition 1 *Suppose* $\{\{e\}\} \succ \{\{h\}\} \succ \{\{h, \neg h\}\}$. *Then*

$$\{\{\neg h, e\},\ \{h, \neg h, e\}\} \succsim \{\{\neg h\},\ \{h, \neg h\}\}.$$

Thus, adding the virtuous option is always weakly desired. Intuitively, $\{\neg h\}$ tempts the agent in $\{\{\neg h\}, \{h, \neg h\}\}$ (because it offers the tempting $\neg h$ without guilt) but adding virtuous alternative both introduces guilt and possibly enhances more virtuous consumption, yielding a menu $\{\neg h, e\}$ that is less tempting than $\{\neg h\}$. In fact, now temptation preferences will be aligned with normative preference, with both preferring $\{h, \neg h, e\}$ over $\{\neg h, e\}$, since guilt cannot be avoided and temptation was arising only because of guilt-avoidance. It is worth noting that even if e is never chosen, the agent is better off for having reduced the temptation she faces in the interim periods. In GP's model, adding e is valuable if and only if it is chosen.

One can similarly analyze an agent's use of resolutions that cause him to self-impose penalties or rewards to different choices. For instance, starting

with the menu $\{\{\neg h\},\ \{h, \neg h\}\}$, the agent can make a resolution to eat healthy and transform his menu into $\{\{\neg h - c\},\ \{h, \neg h - c\}\}$ where $\neg h$ is now associated with a penalty c of breaking his resolution.

5.1 Guilt Avoidance

There is evidence in psychology that agents may "downshift" in order to avoid having to live up to their normative objectives. For instance, they may avoid important tasks in which they do not expect to perform well (Ferrari et al [22]) or enter less ambitious educational programs (Enns et al [21]). In our model such behavior is a property of interim choice, which maximizes

$$U(a) + V(b) = \max_a(u + v - \max_a v) + k \max_a(v + u - \max_a u).$$

The guilt term $u - \max_a u$ gives rise to a guilt-avoidance motive for interim temptation preference, and would push the agent to avoid menus that contain options normatively better than what he anticipates choosing. In the example in the previous section, the agent would be tempted to avoid keeping healthy foods at home if he anticipates eating less healthy instead.

Our model suggests that plans will tend to be virtuous when they are made sufficiently in advance of the time of final consumption (as in the previous section), and as this time gets closer the agent may seek opportunities of avoiding having to face them particularly if he expects to fail.

6 Application: Moral Hypocrisy

Various theories in social psychology try to explain why moral people behave immorally. It may be because moral values have not been taught well enough for them to withstand temptation, or it may be because people's capacity for selective perception and rationalization enables moral disengagement in particular situations (Bandura [7, 8]). A relatively recent theory adds a third possible explanation: the key motivation for moral behavior may come from peoples' desire to *appear* moral. Indeed, studies find that if agents can avoid having to behave morally and pursue self-interest instead, they do. This behavior is referred to as *moral hypocrisy* (Batson et al [5]). Some of the evidence is as follows:

- In seminal work in psychology, Batson et al [5] conducted an experiment where subjects were required to allocate two tasks between themselves

and a partner. The "desirable" task came with the opportunity to win money while the "undesirable" task was dull and payed nothing. They were told that their partner would not be informed that they were allowed to assign the tasks. Subjects were given the opportunity to *privately* flip a fair coin in order to help them make the allocation decision. The researchers found that the coin flippers allocated the desirable task to themselves 90% of the time, even though they privately flipped a fair coin. In a follow-up questionnaire, 75% of subjects who assigned themselves the desirable task indicated that they believed that the morally correct thing to do was to assign it to the partner. The results have been replicated multiple times in the psychology literature.

- In economics, moral hypocrisy has been demonstrated in dictator game experiments. A 'dictator' is asked to allocate an endowment between himself and a passive 'recipient'. Lazear et al [34] find that while (experienced) dictators share with recipients on average 20-30% of their \$10 endowment when playing the dictator game, 50% of them exit the game with the full endowment when given the option. Dana et al [12] find that a significant proportion exit the game even at a cost. Hammam et al [28] find that, when given the option, dictators delegate the endowment allocation decision to agents who tend to be more favorable toward the dictator. Dana et al [13] demonstrate that agents exploit moral wiggle room (such as uncertainty about the outcome of an action for others) in order to behave selfishly. Thus, dictators that share, thereby appearing to possess a sense of morals, also exploit opportunities that increases their share of the pie.
- Anecdotal evidence includes the observation that people who contribute to beggars when they encounter them may also cross the street to avoid encountering them.

7 Application: Dictator Games

While there are several models that explain findings from experiments on dictator games (such as [18, 40]) in this section we show that a specialized version of our model can also unify the existing evidence on dictator games. A generic alternative available to a dictator is an allocation $x = (x_1, x_2)$ where x_1 denotes her own consumption and x_2 denotes the recipient's consumption. Consumption may potentially be lotteries. A dictator playing the game faces a menu a of possible allocations. If the endowment is $(M, 0)$ and the dictator

can share any part of her endowment, then we denote her menu by

$$dg = \{(M - s, s) : 0 \leq s \leq M\}.$$

The option to exit the game with $\$M$ is the singleton menu:

$$e = \{(M, 0)\}.$$

If the dictator is offered a choice of either playing the game or exiting, she faces the menu of menus:

$$A = \{dg, e\}.$$

Assume that a dictator's normative preference is for an equal division of the pie, and that the temptation preference is to maximize own material payoff. Normative utility is u and temptation utility is λv, where λ is a scalar that parametrizes the intensity of temptation. Note that the dictator's choice from any menu can be written as

$$\begin{aligned}\mathcal{C}(a) &= \arg\max_{x \in a} \left\{ \left(u(x) - \left[\max_{z \in a} u(z) - u(x) \right] \right) + \lambda \left(v(x) - \left[\max_{y \in a} v(y) - v(x) \right] \right) \right\} \\ &= \arg\max_{x \in a} \{u(x) + \lambda v(x)\}.\end{aligned}$$

Evidently, greater intensity λ of temptation is associated with lower self-control. Finally, the dictator's choice over menus is given by

$$\mathcal{C}_1(A) = \arg\max_{x \in a} \{U(a) + V(a)\}.$$

where the parameter κ in the model has been set to 1 for simplicity.

We assume that $u(M - s, s)$ is a hump-shaped function of s, with the maxima at $s = \frac{M}{2}$. That is, the agent normatively desires an even split. Assume also that $v(M - s, s)$ is just a function of own payoff, and therefore, v is strictly decreasing in s. Both functions are twice differentiable and concave in s. An immediate observation is:

Proposition 2 *For a dictator with temptation intensity λ, denote the choice by*

$$\mathcal{C}(dg) = \{(M - s_\lambda, s_\lambda)\}.$$

Then s_λ is decreasing in λ.

We show that our model can accommodate some basic facts.

Dictators tend to share though they share less than 50%: In our model, since $\mathcal{C}(a)$ maximizes $u + \lambda v$ the agent's choice is a compromise between her normative desire to share 50% and her temptation to share nothing. Thus, the agent will tend to share but not as much as 50%.

Dictator exhibit a strict preference for exit: Dana et al [12] find that over 25% of dictators exit at a cost (specifically, they exit with \$9 when the pie is worth \$10). Broberg et al [11] find that 64% of their dictators exhibit willingness to exit the game for as little as 82% of the pie.

In our model, since $\mathcal{C}_1(A)$ maximizes $U+V$ and since V maximizes temptation utility less guilt costs, the agent may exhibit a strict preference for committing to a selfish option even at a price – commitment to a selfish option implies that there will be no ex post guilt from consuming that option. If today's guilt from choosing this option is not overwhelming, the agent will commit to the selfish option.

Proposition 3 *There exists λ^*, λ_* s.t. for all $\lambda_* \leq \lambda \leq \lambda^*$,*

$$\mathcal{C}(dg) = \{(M-s, s)\},\ s > 0,\ \textit{and}\ \mathcal{C}_1(\{dg, e\}) = \{e\}$$

The proposition tells us that there are values of λ for which the agent may share when playing the dictator game but also strictly prefer to exit if given the opportunity. At these values, there is enough self-control to share when playing the dictator game, but not enough self-control for guilt to be relatively unimportant and thus for normative preference to overcome the temptation desire to exit.

Dictators that share less also care less about exit: Dana et al [12] find that of those subjects who (under anonymity) offer nothing to the receiver, only 1 of 24 subjects took the exit option. Broberg et al [11] find that subjects who offer nothing also value exit less than those subjects who offer positive amounts.

In our model if $\mathcal{C}(a)$ offers nothing then that is indicative of high λ. That is, the intensity of a temptation to be selfish is high. For the same agents, $\mathcal{C}_1(A)$ would care less for commitment: observe that V maximizes temptation utility less guilt costs and that higher λ implies a *relatively lower importance*

of guilt costs. Since the preference for commitment comes from guilt costs, this implies a lower desire to exit at a price.

The reservation price c for the exit option is defined implicitly by

$$U(\{(M-c,0)\}) + V(\{(M-c,0)\}) = U(a) + V(a).$$

Higher the value of c, the less willing an agent is to exit at a given price.

Proposition 4 *There exists* λ^{**} *s.t.* $\frac{\partial c}{\partial \lambda} < 0$ *for all* $\lambda \geq \lambda^{**}$.

While dictators with lower self-control will share less with recipients, the proposition tells us that for sufficiently low self-control, dictators will not only share nothing but will also not value commitment highly.

The cut off point λ^{**} is strictly less that the λ^{*} in Proposition 3. Agents with temptation intensity $\lambda^{**} \leq \lambda \leq \lambda^{*}$ share with recipients, but the more they share, the higher their temptation preference for exiting.

8 Application: Strategic Ignorance

Ehrich and Irwin [19] demonstrate that consumers avoid obtaining information on the ethicality of products, such as whether the wood used in a piece of furniture comes from endangered rain forests, or whether a cell phone was made by a company with overseas factories that employ child labor. The researchers find that such "willful ignorance" manifested most strongly among those people who had claimed to care most about the ethical issue at hand. That is, those that are likely to experience guilt more intensely also exhibit a greater tendency to avoid information that could cause guilt. Relatedly, Dana et al [13] show that in the context of dictator game experiments, subjects seek 'moral wiggle room' when possible.

While anomalous preference for information has been addressed in the Non-Expected Utility literature (Kreps and Porteus [32], Alaoui [1]), we show that our model can accommodate such phenomena.

8.1 General Context

To show that this is possible in our model, suppose that a piece of furniture is either made locally l or could be manufactured abroad, either in a country where the wood is not obtained by cutting down rain forests f^{*} or one where

it may well be f_*. Assume that the imported furniture is higher quality than local furniture, the agent is ethically opposed to f_* but that he is neverthless tempted by it. More specifically assume:

$$v(l) \leq v(f^*) < v(f_*) \text{ and } u(f_*) < u(l) < u(f^*).$$

The agent could either go to a local furniture store $\{l\}$ or to one that sells $\{l, f^*\}$ or to one that sells $\{l, f_*\}$. If the agent does not inquire about where the imported furniture is manufactured he faces the menu $\{l, \alpha f^* + (1-\alpha) f_*\}$ whenever he goes to any shop that sells local and imported funiture – here a choice of imported furniture is a lottery $\alpha f^* + (1-\alpha) f_*$ with the obvious interpretation.

The choice to obtain information about the origin of the furniture depends on the preference over the following two menus of menus:

$$\{\{l\},\ \{l, f^*\},\ \{l, f_*\}\} \text{ versus } \{\{l\}, \{l, \alpha f^* + (1-\alpha) f_*\}\}.$$

We prove general results in Appendix F but present a concrete example here. Let $v(l) = v(f^*) = 0$ and $v(f_*) = 1$, while $u(f_*) = 0$, $u(l) = 0.25$ and $u(f^*) = m > 0.25$. Let α be sufficiently close to 0 so that l is normatively better than $\alpha f^* + (1-\alpha) f_*$. With these values, the agent never chooses l in any non-singleton menu containing it, and in particular experiences guilt in menus $\{l, f_*\}$ and $\{l, \alpha f^* + (1-\alpha) f_*\}$. We find that there are instances where ignorance can be desirable, and others where it may be harmful.

Proposition 5 *If* $m > 0.75$ *then*

$$\{\{l\},\ \{l, f^*\},\ \{l, f_*\}\} \succ \{\{l\}, \{l, \alpha f^* + (1-\alpha) f_*\}\}.$$

If $0.75 \geq m > 0.25$ *then*

$$\{\{l\}, \{l, \alpha f^* + (1-\alpha) f_*\}\} \succ \{\{l\},\ \{l, f^*\},\ \{l, f_*\}\}.$$

Ignorance can exacerbate temptation and therefore may be undesirable. The choice of f_* from $\{l, f_*\}$ comes at the guilt cost of not choosing l. The choice of $\alpha f^* + (1-\alpha) f_*$ from $\{l, \alpha f^* + (1-\alpha) f_*\}$ has lower temptation value, but also lower guilt cost relative to l (because there is a chance it is the normatively superior f^*). When m is "large" the reduction in guilt outweighs the lower temptation value, and so $\{l, \alpha f^* + (1-\alpha) f_*\}$ is more tempting than $\{l, f_*\}$, ie, $V(\{l, \alpha f^* + (1-\alpha) f_*\}) > V(\{l, f_*\})$. Consequently

we could have a scenario where the agent would exert self-control in the interim stage if he had information, and would submit to temptation if he did not.

On the other hand ignorance may be desirable if knowledge, rather than ignorance, worsens temptation. In such a case we could have a scenario where the agent always submits to temptation both ex post and in the interim. Then with information the agent selects $\{l, f_*\}$ in the interim period and f_* ex post, whereas without information he selects $\{l, \alpha f^* + (1-\alpha) f_*\}$ in the interim period and $\alpha f^* + (1-\alpha) f_*$ ex post. The latter is normatively better in that the agent probabilistically improves his outcome via ignorance.

Next we consider the case of partial information, where the agent only inquires which store carries f^*, in which case we need to study the preference between

$$\{\{l\},\ \{l, f^*\},\ \{l, f_*\}\} \quad \text{versus} \quad \{\{l\},\ \{l, f^*\},\ \{l, \alpha f^* + (1-\alpha) f_*\}\}.$$

In the same example we have

Proposition 6 *If* $m \geq 1$ *then*

$$\{\{l\}, \{l, f^*\}, \{l, f_*\}\} \succ \{\{l\}, \{l, f^*\}, \{l, \alpha f^* + (1-\alpha) f_*\}\}$$

If $0.75 \geq m > 0.25$ *then*

$$\{\{l\}, \{l, f^*\}, \{l, \alpha f^* + (1-\alpha) f_*\}\} \succ \{\{l\}, \{l, f^*\}, \{l, f_*\}\}.$$

The cases are interpreted similarly as in the previous proposition. When m is large then ignorance enhances temptation due to lower guilt, and in this case full information is strictly better than ignorance of the stores selling the normatively-inferior commodity. When m is smaller, then knowledge can worsen temptation.

8.2 Dictator Game Context

In Dana et al [13], subjects had the option of playing a random version of the dictator game:

$$a = \{(6, \ell), (5, \ell)\},$$

where (x, ℓ) reflects an allocation yielding x to the dictator and ℓ to the recipient, and moreover, $\ell = (\frac{1}{2}, 5; \frac{1}{2}, 1)$ is the lottery yielding \$5 or \$1 to the

recipient with even probabilities. Dictators also had the option of resolving the uncertainty, and being left with one of the following dictator games with 50% probability:

$$a_1 = \{(6,5),(5,1)\} \text{ or } a_2 = \{(6,1),(5,5)\}.$$

Observe that in a_1 the selfish choice is also fairest choice. Approximately half (44%) of the subjects preferred not to reveal the information, and of these the majority (86%) chose the selfish option.

Our model would explain this in terms of interim choice of menus in the following way. When faced with a the agent may guiltlessly choose to be selfish as her choice does not affect the distribution of the recipient's outcome. Acting selfishly in a_1 would also be guiltless but not in a_2. Therefore, when self-control is low (so that she expects to act selfishly in any menu), not revealing the information permits a selfish action without guilt, whereas revealing the information would lead to a selfish action with strictly positive *expected* guilt. This will be reflected in first period choice if the first period guilt from such a choice is not overwhleming. Denote by $(\frac{1}{2}, a_1; \frac{1}{2}, a_2)$ then lottery over a_1 and a_2.

Proposition 7 *There exists* λ^{***} *s.t.* $\mathcal{C}_1(\{a, (\frac{1}{2}, a_1; \frac{1}{2}, a_2)\}) = \{a\}$ *for all* $\lambda \geq \lambda^{***}$.

9 Related Models

The GP [26] model has a two period time line such that in period 1 the agent ranks menus and in period 2 chooses out of a menu. Their model adopts a preference $\succsim^1$ over menus as the primitive and axiomatize a representation of the form

$$U(a) = \max_{x \in a}[u(x) - \left(\max_{y \in a} v(y) - v(x)\right)].$$

While period 1 choice is unmodelled, there are results in the literature (Noor [37]) that show how their model can be augmented so as to provide a joint representation for period 1 and period 2 choice.

Notions closely related to guilt have been studied in GP-type models in the decision theory literature. Dillenberger and Sadowski [18] specialize to the dictator game setting, and study the shame associated with behaving

selfishly when being observed by the recipient. The model is a discrete version of the GP representation, where u is interpreted as the agent's private preferences – interpreted as selfish – and v reflects the perceived social norm. If the agent behaves selfishly relative to the norm then she experiences the cost of shame, in the same maner that the GP agent experiences the cost of self-control when choice follows u rather than v. Saito [40] extends this model (using the same domain as GP but still specializing to the dictator game setting) to include other psychological features such as pride. Both papers adhere to the social image theory of social preferences.

Kopylov [30] studies the negative emotions experienced by a perfectionist who is unable to meet her perfectionist standards. A special case of his model that is relevant here is the extention of GP given by:

$$W(a) = \max_{x \in a} \{u(x) - \left(\max_{y \in a} v(y) - v(x) \right) + k \left(\max_{y \in a} u(y) - u(x) \right) \}.$$

Here, u represents the agent's perfectionist standards and v represents desires, and the new term in the representation is the cost (guilt, anger, disappointment) of deviating from her perfectionist standards.[10]

Though not intended as such by the authors, these models can be reinterpreted as models of guilt in obvious ways. To see how these reinterpreted models are different from our model, observe that the preference $\succsim^1$ over menus in both models exhibits a guilt-avoidance motive. But these reinterpreted models would be consistent with the example discussed under observation N in Section 2.1: they would exhibit,

$$\{y - \varepsilon\} \succ^1 \{x, y\},$$

where $y - \varepsilon$ is a normatively worse alternative than y and x is the good alternative. The reinterpreted models would attribute such behavior to the agent's ex-ante (and thus normative) preference. However we have argued in this paper that a model of guilt should attribute it to temptation (Section 2).

[10] Observe that interim choice of menus in our model maximizes the utility

$$(U + V)(a) = \max_{x \in a} \{u(x) + \kappa v(x) - [\max_{y \in a} v(y) - v(x)] + \kappa [\max_{y \in a} u(y) - u(x)]\}.$$

This representation is similar to Kopylov [30] except that preferences over *singleton* menus in Kopylov [30] is represented by normative utility u rather than a compromise of normative and temptation utility $u + \kappa v$ as in our model.

Within the menus literature but outside the temptation literature, Sarver [41] introduces a model where a preference for commitment arises due to regret-aversion. In the ex post period the agent is uncertain about her utility $u \in \mathcal{U}$, which is drawn from some distribution π. She must make a choice from a menu prior to the realization of u. However, for any alternative x that she chooses from menu a and any realized u, she regrets not having chosen the maximizer of u (captured by the difference $max_{y \in a} u(y) - u(x)$). Her choice maximizes expected utility net of regret. Her ex ante preference is represented by the value function of the problem, $W(a) = \max_{x \in a} \int_{\mathcal{U}} [u(x) - K\left(\max_{y \in a} u(y) - u(x)\right) d\pi(u)$. The agent may ex ante prefer a smaller menu in order to avoid regret. The model does not readily admit an interpretation in terms of guilt. But similar to guilt, the models of both Sarver [41] and Kopylov [30] would give rise to an infinite regress, although this is not accounted for in these papers. The models of Dillenberger and Sadowski [18] and Saito [40] do not give rise to an infinite regress since period 0 is qualitatively different from period 1: the ex post choice is observed by some observer while ex ante choice is hypothesized to be unobserved.

Finally, we note that this paper does not overlap with the work on guilt by Battigalli and Dufwenberg [6]. These authors define guilt by the utility loss experienced when behavior falls short of others' expectations. In contrast, in our model, guilt is the consequence of behavior falling short of an abstract normative preferences. An interesting hypothesis for consideration in future research is that normative preferences could depend on expectations. Another difference is that Battigalli and Dufwenberg [6] focus on strategic settings whereas we focus on a single decision maker.

A Appendix: Infinite Horizon version

In this section we formulate our framework as in Noor [37], and present the main idea as to how U can be elicited from choice when choice is subject to temptation and guilt in every period.

Suppose time is given by $1, 2, ...$ For our domain, adopt the infinite horizon choice problems $\mathcal{M}$ in GP [27], which they show to be homeomorphic to $\mathcal{K}(\Delta(C \times \mathcal{M}))$. Thus, a menu $x \in \mathcal{M}$ can be treated as a compact set of alternatives, where each alternative is a lottery over pairs (c, y) that yields immediate consumption $c \in C$ and another menu $y \in \mathcal{M}$ in the next period.

In this setting the choice in any period is modelled as the correspondence $C : \mathcal{M} \rightsquigarrow \Delta(C \times \mathcal{M})$ defined by

$$C(x) = argmax_{\mu \in x}[U(\mu) - \left[\max_{\eta \in x} V(\eta) - V(\mu)\right] + V(\mu) - \left[\max_{\eta \in a} U(\eta) - V(\mu)\right]]$$

where

$$U(\mu) = \int_{C \times \mathcal{M}} u(c) + \delta W(x) d\mu(c, x) \quad \text{s.t.} \quad W(x) = \max_{\mu \in x}[U(\mu) - \left[\max_{\eta \in x} V(\eta) - V(\mu)\right]]$$

and

$$V(\mu) = \int_{C \times \mathcal{M}} u(\mu) + \kappa G(x) d\mu(c, x) \quad \text{s.t.} \quad G(x) = \max_{\mu \in x}[V(\mu) - \left[\max_{\eta \in a} U(\eta) - V(\mu)\right]]$$

for some continuous $u, v : C \to \mathbb{R}$ and $0 < \kappa < \delta < 1$. Thus, choice of a lottery $\mu \in x$ maximizes the compromise between normative and temptation utilities net of self-control cost and guilt. The agent's normative perspective U on (c, x) relies on the normative utility $u(c)$ from immediate consumption and the δ-discounted normative utility $W(x)$ from the continuation menu. Similarly the temptation perspective V on (c, x) relies on the temptation utility $v(c)$ from immediate consumption and the κ-discounted temptation utility $G(x)$ from the continuation menu. The normative utility $W(x)$ from a menu x maximizes U less self-control cost, as in GP [26]. The temptation utility $G(x)$ maximizes V less guilt-cost as hypothesized in the current paper. The assumption that $\kappa < \delta$ states that if we delay the receipt of a menu then its temptation utility weakens faster than normative utility.

While the choice $C(x)$ from menu x maximizes normative and temptation utilities net of self-control and guilt costs, it is equivalent to just maximizing normative and temptation utility:

$$C(x) = argmax_{\mu \in x}[U(\mu) - \left[\max_{\eta \in x} V(\eta) - V(\mu)\right] + V(\mu) - \left[\max_{\eta \in a} U(\eta) - V(\mu)\right]]$$

$$\implies C(x) = argmax_{\mu \in x}[2U(\mu) + 2V(\mu) - \max_{\eta \in x} V(\eta) - \max_{\eta \in a} U(\eta)]$$

$$\implies C(x) = argmax_{\mu \in x}[U(\mu) + V(\mu)],$$

so the last expression can be taken for simplicity.

We now show how to derive U from choice C. Suppose $0 \in C$ is an alternative such that $u(c) = v(c) = 0$. For any $t > 1$, a t-dated menu is an alternative $x^{+t} \in \Delta(C \times \mathcal{M})$ that is committed to yielding consumption 0 in periods $1, .., t-1$ and then menu x in period t.[11] Define a revealed preference $\succsim^{+t}$ over $\mathcal{M}$ by the choice over t-dated menus and observe that

$$x \succsim^{+t} y \iff x^{+t} \in C(\{x^{+t}, y^{+t}\})$$

$$\iff U(x^{+t}) + V(x^{+t}) \geq U(y^{+t}) + V(y^{+t})$$

$$\iff \delta^t W(x) + \kappa^t G(x) \geq \delta^t W(y) + \kappa^t G(y)$$

$$\iff W(x) + (\frac{\kappa}{\delta})^t G(x) \geq W(y) + (\frac{\kappa}{\delta})^t G(y).$$

Since $\frac{\kappa}{\delta} < 1$,

$$W(x) + (\frac{\kappa}{\delta})^t G(x) \to W(x) \text{ as } t \to \infty.$$

This implies that $\succsim^{+t}$ converges (in an appropriate topology – see Noor [37]) to some preference $\succsim^*$ over $\mathcal{M}$ that represents the normative utility over menus W.[12] That is, if our primitive consists of choices C that admit the above representation, then although C is subject to temptation and guilt, we can uniquely identify the agent's normative perspective from C by taking a limit of revealed preferences over delayed menus. Uniqueness properties of the representation can be established using this.

The model in the current paper circumvents the technical details by presuming simply that the period 0 preference over menus is $\succsim^*$. It does so on the grounds that revealed preference foundations for such a primitive can be obtained as in Noor [37]. In an axiomatization of the model, conditions will be required so that $\succsim^*$ satisfies the axioms in this paper.

B Appendix: Proof of Theorem 1

Lemma 1 *There exist $\kappa, k_2, k_3 \in \mathbb{R}$, such that for all $a \in \mathcal{M}_1$,*

$$V(a) = \kappa \max_a (u + v) + k_2 \max_a v + k_3 \max_a u \tag{4}$$

[11] Define this inductively by setting $x^{+1} = x$ and $x^{+t} = \{(0, x^{+(t-1)})\}$ for each $t > 1$.

[12] Although $\delta^t W(x) + \kappa^t G(x) \to 0$, the preference represented by $\delta^t W(x) + \kappa^t G(x)$ does not converge the preference represented by 0.

Proof. We first show that

$$\max_a u = \max_b u \text{ and } \max_a (u+v) = \max_b (u+v) \text{ and } \max_a v = \max_b v \implies V(a) = V(b). \tag{5}$$

This is proved as in Kopylov and Noor [31]. The argument is then completed by appealing to Harsanyi's aggregation theorem.

Suppose the above equalities hold. Note that $U(a) = U(b)$. Suppose by way of contradiction that $V(a) < V(b)$. Then $(U+V)(a) < (U+V)(b)$. Since $\succsim$ is nondegenerate, there exist x^*, y^* such that $U(\{x^*\}) > U(\{x^*, y^*\}) > U(\{y^*\})$. Let $a^* = \varepsilon\{y^*\} + (1-\varepsilon)a$ and $b^* = \varepsilon\{x^*\} + (1-\varepsilon)b$ and take $\varepsilon > 0$ s.t.

$$(U+V)(a^*) < (U+V)(b^*). \tag{6}$$

Such ε exists by continuity of U, V. Note also that by linearity of U, V and u,

$$U(a^*) > U(a^* \cup b^*) > U(b^*) \tag{7}$$

$$\text{and} \max_{a^*} u < \max_{b^*} u. \tag{8}$$

Given (6), there are two possibilities.

- $(U+V)(a^*) < (U+V)(a^* \cup b^*)$. Combined with (7), $V(a^*) < V(a^* \cup b^*)$. Therefore, $\{a^*\} \succ \{a^*, a^* \cup b^*\}$, which contradicts Temptation Sophistication.
- $(U+V)(b^*) > (U+V)(a^* \cup b^*)$. Combined with (7), $V(b^*) > V(a^* \cup b^*)$. Therefore, $\{a^* \cup b^*\} \succ \{a^* \cup b^*, b^*\}$. However, given (8), this contradicts Guilt-Averse Temptation.

This establishes (5). To prove the lemma, first restrict attention the set of convex interim menus:

$$\mathcal{M}_1^c = \{co(a) : a \in \mathcal{M}_1\},$$

where $co(a)$ denotes the convex hull of a with respect to the mixture operation. The set $\mathcal{M}_1^c$ is a mixture space, and thus the pareto condition (5) yields the desired form for V on $\mathcal{M}_1^c$ by an application of Harsanyi's aggregation theorem (see Border [9]).[13] Extend the representation to all menus in $\mathcal{M}_1$

[13] Harsanyi's theorem delivers the desired form plus a constant, but the constant can be set to zero wlog given the uniqueness properties in Theorem 1.

by exploiting the fact that linearity of V implies the 'Indifference to Timing' condition $V(a) = V(co(a))$ for all $a \in \mathcal{M}_1$ (see Dekel, Lipman and Rusticini [16]). ■

Lemma 2 *$\kappa + k_3 = 0$ and $\kappa + k_2 > 0$.*

Proof. Suppose by way of contradiction that $\kappa + k_3 \neq 0$. Consider two cases:

Case i: $\kappa + k_3 > 0$ or $[\kappa + k_3 < 0$ and $\kappa + k_2 < 0]$

We show that Temptation Consistency must be violated. By the non-degeneracy of $\succsim$ and the fact that u and v are nonconstant and affinely independent, there exist x', y', x'', y'' such that:

$$u(x') = u(y'), v(x') < v(y')$$
$$v(x'') = v(y''), u(x'') > u(y'')$$

By the linearity of u and v, for any $\theta \in (0, 1)$,

$$\frac{u(x'\theta x'') - u(y'\theta y'')}{v(y'\theta y'') - v(x'\theta x'')} = \frac{1-\theta}{\theta} \frac{u(x'') - u(y'')}{v(y') - v(x')} := f(\theta).$$

Observe that $f(\theta)$ ranges between 0 and infinity.

We first show that $V(\{x'\theta x''\}) > V(\{y'\theta y''\})$. If $\kappa + k_3 > 0$, then we can find θ such that $f(\theta) > \frac{\kappa+k_2}{\kappa+k_3}$. But then,

$\frac{u(x'\theta x'')-u(y'\theta y'')}{v(y'\theta y'')-v(x'\theta x'')} = f(\theta) > \frac{\kappa+k_2}{\kappa+k_3}$

$\Longrightarrow (\kappa + k_3)\, u(x'\theta x'') + (\kappa + k_2)\, v(x'\theta x'') > (\kappa + k_3)\, u(y'\theta y'') + (\kappa + k_2)\, v(y'\theta y'')$

$\Longrightarrow V(\{x'\theta x''\}) > V(\{y'\theta y''\})$, given the functional form for V established in the previous lemma. On the other hand, if $[\kappa + k_3 < 0$ and $\kappa + k_2 < 0]$ then we find θ such that $f(\theta) < \frac{\kappa+k_2}{\kappa+k_3}$ and use an analogous argument to show that $V(\{x'\theta x''\}) > V(\{y'\theta y''\})$.

Next, observe that the linearity of u and v implies

$$u(x'\theta x'') = \theta u(x') + (1-\theta)u(x'') > \theta u(y') + (1-\theta)u(y'') = u(y'\theta y''),$$

and similarly, $v(x'\theta x'') < v(y'\theta y'')$. Letting $x := x'\theta x''$ and $y := y'\theta y''$, we therefore have

$$u(x) > u(y),\ v(x) < v(y) \text{ and } V(x) > V(y).$$

However, these inequalities imply $\{\{x\}\} \succ \{\{x, y\}\}$ and $\{\{x\}\} \sim \{\{x\}, \{y\}\}$, which contradicts Temptation Consistency, as desired.

Case ii: $\kappa + k_3 < 0$ and $\kappa + k_2 \geq 0$,
By nondegeneracy, there exists $x, y \in \triangle(Z)$ such that

$$\{\{x\}\} \succ \{\{x, y\}\}.$$

By the representation, $u(x) > u(y)$ and $v(x) < v(y)$. Moreover, by the form for V,

$$\begin{aligned} V(\{x\}) &= (\kappa + k_3) u(x) + (\kappa + k_2) v(x) \\ &< (\kappa + k_3) u(y) + (\kappa + k_2) v(y) \\ &= V(\{y\}). \end{aligned}$$

But then $\{\{x\}\} \sim \{\{x\}, \{y\}\}$, which contradicts Temptation Consistency.

We have therefore shown that $\kappa + k_3 = 0$. Observe that $V(\{x\}) = (\kappa + k_2) v(x)$. Suppose by way of contradiction that $\kappa + k_2 \leq 0$. By nondegeneracy, there is x^*, y^* such that $u(x^*) > u(y^*)$ and $v(x^*) < v(y^*)$. However, it would then follow that $V(\{x^*\}) > V(\{y^*\})$, and in particular, $\{\{x^*\}\} \succ \{\{x^*, y^*\}\}$ and $\{\{x^*\}\} \sim \{\{x^*\}, \{y^*\}\}$, contradicting Temptation Consistency. Thus, $\kappa + k_2 > 0$. This completes the proof. ■

Lemma 3 $k_2 = 0$.

Proof. By nondegeneracy, there is x, y such that $\{\{x\}\} \succ \{\{x, y\}\} \succ \{\{y\}\}$. By the representation, $u(x) > u(y)$, $v(x) < v(y)$ and $u(x) + v(x) > u(y) + v(y)$. However, by Temptation Sophistication,

$\{\{x, y\}\} \succ \{\{y\}\}$
$\Longrightarrow \{\{x\}\} \sim \{\{x\}, \{x, y\}\}$
$\Longrightarrow V(\{x, y\}) \leq V(\{x\})$
$\Longrightarrow \kappa v(x) + k_2 v(y) \leq \kappa v(x) + k_2 v(x)$ by previous lemmas
$\Longrightarrow k_2 v(y) \leq k_2 v(x)$.
Since $v(x) < v(y)$, it follows that $k_2 = 0$, as desired. ■

Lemma 4 $\kappa < 1$ *iff Preference for Early Choice holds.*

Proof. We prove sufficiency. By nondegeneracy and Basic, there exists some consumption $x, y \in C$ s.t.

$$\{\{x\}\} \succ \{\{x, y\}\} \succ \{\{y\}\},$$

and so $v(y) - v(x) > 0$ while

$$W(\{\{x, y\}\}) = u(x) - [v(y) - v(x)].$$

By Temptation Consistency and Preference for Early Choice,

$$\{\{x\}\} \succ \{\{x\}, \{y\}\} \succ \{\{y\}\}$$

and so

$$W(\{\{x\}, \{y\}\}) = u(x) - \kappa [v(y) - v(x)].$$

By Preference for Early Choice and $v(y) - v(x) > 0$, it follows that $\kappa < 1$. ■

Lemma 5 *The representation (u, v, κ) is unique in the sense that (u', v', κ') is the representation for the same preference if and only if, $u' = \alpha u + \beta_1$, $v' = \alpha v + \beta_2$, and $\kappa' = \kappa$.*

Proof. The first two requirements are given in GP. For the third part, if $\kappa' \neq \kappa$, then $V' = \kappa' [\max(u' + v') - \max u'] = \alpha\kappa' [\max(u + v) - \max u] + \alpha\kappa'\beta_2$. According to theorem 1, $V' = \alpha V + \beta_3$. Combine these two equations, $\beta_3 = \alpha\kappa'\beta_2 + \alpha(\kappa' - \kappa)[\max(u + v) - \max u]$, which is not a constant given the nondegeneracy conditions, a contradiction. ■

C Appendix: Proof of Theorems 3 and 4

Suppose $\succsim$ and $\succsim^*$ are a pair of nondegenerate preferences with guilt representations (u, v, κ) and (u^*, v^*, κ^*) respectively.

Lemma 6 *$\succsim$ and $\succsim^*$ are ex post similar if and only if wlog $v = v^*$ and $u = \lambda u^*$ for some $\lambda > 0$.*

Proof. The claim '$\Longleftarrow$' is trivial, so consider '$\Longrightarrow$'.The claim that $u = \lambda u^*$ for some $\lambda > 0$ is obvious. Suppose by way of contradiction that v and v^* are ordinally distinct. Then there is x, y s.t. wlog $v(x) \geq u(y)$ and $v^*(x) \leq v^*(y)$ with one strict inequality. Since nondegeneracy implies that v and v^* are nonconstant, we can assume $v(x) > v(y)$ and $v^*(x) < v^*(y)$ wlog. If $u(x) < u(y)$ then ex post similarity implies $\{\{y\}\} \succ^* \{\{y, x\}\}$ and thus $v^*(x) > v^*(y)$, a contradiction. Thus $u(x) \geq u(y)$.

By nondegeneracy and ex post similarity,

$$\{\{w\}\} \succ \{\{w,z\}\} \text{ and } \{\{w\}\} \succ^* \{\{w,z\}\}$$

for some w, z. Observe that by the representation, for all α,

$$\{\{\alpha w + (1-\alpha)\}\, x\} \succ^* \{\{\alpha w + (1-\alpha)x, \alpha z + (1-\alpha)y\}\}.$$

However, there exists α such that

$$\{\{\alpha w + (1-\alpha)\}\, x\} \not\succ \{\{\alpha w + (1-\alpha)x, \alpha z + (1-\alpha)y\}\},$$

contradicting ex post similarity. Therefore, v and v^* are ordinally equivalent. Since both are linear, they are cardinally equivalent. By the uniqueness result in Theorem 1, we can take $v = v^*$ wlog by redefining λ if necessary. ■

Lemma 7 *$\succsim$ is more guilt-prone than $\succsim^*$ if and only if $\lambda \geq 1$.*

Proof. $\Longrightarrow$: Suppose by way of contradiction that $\lambda < 1$. By the nondegeneracy of $\succsim$ and the fact that u and v are nonconstant and affinely independent, there exist x', y', x'', y'' such that:

$$u(x') = u(y'), v(x') < v(y')$$
$$u(x'') > u(y''), v(x'') = v(y'')$$

By the linearity of u and v, for any $\theta \in (0,1)$,

$$\frac{u(x'\theta x'') - u(y'\theta y'')}{v(y'\theta y'') - v(x'\theta x'')} = \frac{1-\theta}{\theta}\frac{u(x'') - u(y'')}{v(y') - v(x')} := f(\theta)$$

Choose θ such that

$$\lambda < f(\theta) < 1.$$

Let $x = x'\theta x''$, $y = y'\theta y''$ and $z = z'\theta z''$. Then,

$$f(\theta) < 1 \Longrightarrow \frac{u(x) - u(y)}{v(y) - v(x)} < 1 \Longrightarrow u(y) + v(y) > u(x) + v(x), \tag{9}$$

$$f(\theta) > \lambda \Longrightarrow \frac{u(x) - u(y)}{v(y) - v(x)} > \lambda \Longrightarrow \frac{u(y)}{\lambda} + v(y) < \frac{u(x)}{\lambda} + v(x), \tag{10}$$

$$u(x) > u(y), v(x) < v(y) \tag{11}$$

Let $a=\{x\}$, $b=\{y\}$. Then $a\cup b=\{x,y\}$. By (11),

$$\{a\}\succ\{a\cup b\},\{a\}\succ^{*}\{a\cup b\} \tag{12}$$

By (9),

$$\left.\begin{array}{c} V(a\cup b)=\kappa\left[u(y)+v(y)-u(x)\right]>\kappa v(x) \\ V(a)=\kappa\left[u(x)+v(x)-u(x)\right]=\kappa v(x) \end{array}\right\}$$
$$\Longrightarrow V(a\cup b)>V(a)$$

This implies, together with (12),$\{a\}\succ\{a,a\cup b\}$.
On the other hand, by (10),

$$V^{*}(a\cup b)=\kappa^{*}\left[\tfrac{u(x)}{\lambda}+v(x)-\tfrac{u(x)}{\lambda}\right]=\kappa^{*}v(x)$$
$$V^{*}(a)=\kappa^{*}\left[\tfrac{u(x)}{\lambda}+v(x)-\tfrac{u(x)}{\lambda}\right]=\kappa^{*}v(x)$$
$$\Longrightarrow V^{*}(a\cup b)=V^{*}(a)$$

Together with equation (12), $\{a\}\sim^{*}\{a,a\cup b\}$, which contradicts with that $\succsim$ is more guilt-prone then $\succsim^{*}$.

$\Longleftarrow$: If $\lambda=1$ then the proof is trivial. So suppose $\lambda>1$.The preference $\{a\}\succ\{a,a\cup b\}$ implies,

$$U(a)>U(a\cup b)\Longrightarrow u(x)>\max_{a\cup b}\left[u+\left(v-\max_{a\cup b}v\right)\right]$$
$$V(a)<V(a\cup b)\Longrightarrow v(x)<\max_{a\cup b}\left[v+\left(u-\max_{a\cup b}u\right)\right]$$

By these inequalities and by definition of V^{*},

$$\begin{aligned} V^{*}(a\cup b)&=\kappa^{*}\max_{a\cup b}\left[v+\frac{1}{\lambda}\left(u-\max_{a\cup b}u\right)\right] \\ &\geq\kappa^{*}\max_{a\cup b}\left[v+\left(u-\max_{a\cup b}u\right)\right] \\ &>\kappa^{*}v(x)=V^{*}(a) \end{aligned}$$

On the other hand, by the definition of U^{*},

$$\begin{aligned} U^{*}(a\cup b)&=\max_{a\cup b}\frac{1}{\lambda}\left[u+\lambda\left(v-\max_{a\cup b}v\right)\right] \\ &\leq\max_{a\cup b}\frac{1}{\lambda}\left[u+\left(v-\max_{a\cup b}v\right)\right] \\ &<\frac{1}{\lambda}u(x)=U^{*}(a) \end{aligned}$$

Therefore, $\{a\} \succ^* \{a, a \cup b\}$. ■

Lemma 8 *$\lambda \geq 1$ if and only if $\succsim$ has more ex post self-control than $\succsim^*$.*

Proof. $\Longrightarrow$: If $\lambda = 1$ then the proof is trivial. So suppose $\lambda > 1$ and let $\{\{x\}\} \succ^* \{\{x, y\}\} \succ^* \{\{y\}\}$. Then $u^*(x) > u^*(y)$, $v^*(x) < v^*(y)$ and $u^*(x) + v^*(x) > u(y) + v(y)$. By the previous lemma, $u(x) > u(y)$ and $v(x) < v(y)$. Moreover by $\lambda > 1$,

$$u(x)+v(x) = (\lambda - 1)\, u^*(x)+u^*(x)+v^*(x) > (\lambda - 1)\, u^*(y)+u^*(y)+v^*(y) = u(y)+v(y).$$

Therefore, $\{\{x\}\} \succ \{\{x, y\}\} \succ \{\{y\}\}$.

$\Longleftarrow$: Suppose by way of contradiction that $\lambda < 1$. By the construction used in the previous lemmas, we can find x, y such that $u(x) > u(y)$, $v(x) < v(y)$, $u(x) + v(x) < u(y) + v(y)$, but $\frac{u(x)}{\lambda} + v(x) > \frac{u(y)}{\lambda} + v(y)$ (let $f(\theta) \in \left(1, \frac{1}{\lambda}\right)$). That is,

$$\{\{x\}\} \succ \{\{x, y\}\} \sim \{\{y\}\}$$
$$\{\{x\}\} \succ^* \{\{x, y\}\} \succ^* \{\{y\}\},$$

a contradiction. ■

D Appendix: Proof of Proposition 1

Compute that

$(U + V)(\{\neg h, e\}) = (1 + k) \max_{\{\neg h, e\}} u + v - v(\neg h) - ku(e)$, and $(U + V)(\{h, \neg h, e\}) = (1+k) \max_{\{h,\neg h,e\}}(u+v) - v(\neg h) - ku(e)$. So $(U+V)(\{h, \neg h, e\}) \geq (U+V)(\{\neg h, e\})$ iff $(1+k) \max_{\{h,\neg h,e\}}(u+v) - (1+k) \max_{\{\neg h,e\}}(u+v) \geq 0$. The latter inequality always holds therefore the agent always chooses $\{h, \neg h, e\}$ from $\{\{\neg h, e\}, \{h, \neg h, e\}\}$.

Similarly, compute $(U + V)(\{\neg h\}) = (1 + k) \max_{\{\neg h\}}(u + v) - v(\neg h) - ku(\neg h)$, and $(U + V)(\{h, \neg h\}) = (1 + k) \max_{\{h,\neg h\}}(u + v) - v(\neg h) - ku(h)$. So $(U + V)(\{h, \neg h\}) \geq (U + V)(\{\neg h\})$ iff $(1 + k) \max_{\{h,\neg h\}}(u + v) - (1 + k) \max_{\{\neg h\}}(u+v) \geq k[u(h) - u(\neg h)]$. Thus either choice from $\{\{\neg h\}, \{h, \neg h\}\}$ is possible.

It is clear that $V(\{\neg h\}) > V(\{h, \neg h\})$. Since $U(\{h, \neg h\}) \geq U(\{\neg h\})$, we therefore have

$$U(\{h, \neg h\}) \geq W\{\{\neg h\}, \{h, \neg h\}\}.$$

But also, $V(\{h, \neg h, e\}) \geq V(\{\neg h, e\})$, since $V(\{\neg h, e\}) = \max_{\{\neg h, e\}}(u+v) - u(e)$ and $V(\{h, \neg h, e\}) = \max_{\{h, \neg h, e\}}(u+v) - u(e)$. That is, the presence of virtuousness eliminates temptation to commit to indulgence and brings temptation preference in line with normative preference. Thus,

$W\{\{\neg h, e\}, \{h, \neg h, e\}\}$
$= U(\{h, \neg h, e\})$
$= \max_{\{h, \neg h, e\}}(u+v) - v(\neg h)$
$\geq \max_{\{h, \neg h\}}(u+v) - v(\neg h)$
$= U(\{h, \neg h\}) \geq W\{\{\neg h\}, \{h, \neg h\}\}$, as desired. ■

E Appendix: Proofs of Propositions 2-4 and 7

Proof of Prop 2. Due to the hump-shape of normative utility $u(M-s, s)$,

$$\frac{du}{ds} = u_2 - u_1 \begin{cases} < 0 & \frac{M}{2} < s < 1 \\ > 0 & 0 < s < \frac{M}{2} \end{cases}$$
$$\frac{d^2u}{d^2s} = u_{22} - 2u_{12} + u_{11} < 0 \quad s \in [0, M]$$

Similarly, for temptation utility $v(M-s, s)$, for $s \in [0, M]$,

$$\frac{dv}{ds} = v_2 - v_1 < 0$$
$$\frac{d^2v}{d^2s} = v_{22} - 2v_{12} + v_{11} < 0$$

Dictators choose s_λ to maximize $u(M-s, s) + v(M-s, s)$. Since $\frac{du}{ds}, \frac{dv}{ds} < 0$ for any $s > \frac{M}{2}$, it must be that $s_\lambda \leq \frac{M}{2}$. That is, both normative and temptation utilities are decreasing for $s > \frac{M}{2}$, thus optimal choice will be at a level of sharing less than $\frac{M}{2}$.

Let λ_0 be the threshold temptation intensity such that,

$$\frac{du}{ds} + \lambda_0 \frac{dv}{ds}\bigg|_{s=0} = 0. \tag{13}$$

That is, at λ_0 the agent chooses exactly not to share. Given $\frac{d^2u}{d^2s} + \lambda \frac{d^2v}{d^2s} < 0$, it must be that whenever $\lambda \geq \lambda_0$, then

$\frac{du}{ds} + \lambda \frac{dv}{ds}\big|_{s=0} \leq 0$
$\implies \frac{du}{ds} + \lambda \frac{dv}{ds}\big|_{s \in \left(0, \frac{M}{2}\right)} < 0$
$\implies s_\lambda = 0$, that is, the agent continues not to share for any temptation intensity λ higher than λ_0. For λ strictly lower than λ_0, we have $s_\lambda \in \left(0, \frac{M}{2}\right)$, which is determined by the first order condition,

$$FOC : \frac{du}{ds} + \lambda \frac{dv}{ds} = 0.$$

Note that $\lambda = -\left.\frac{du/ds}{dv/ds}\right|_{s=s_\lambda}$. Differentiating wrt s_λ and inverting leads to:

$$\frac{ds_\lambda}{d\lambda} = \left. \frac{(dv/ds)^2}{\underbrace{\left(d^2v/ds^2\right)}_{<0}\underbrace{(du/ds)}_{>0} - \underbrace{\left(d^2u/ds^2\right)}_{<0}\underbrace{(dv/ds)}_{<0}} \right|_{s=s_\lambda} < 0$$

Therefore, we have that as λ increases from 0 to λ_0, s_λ decreases from $\frac{M}{2}$ to 0. ■

Proof of Prop 3. Recall the threshold λ_0 defined in (13). In order for $\mathcal{C}(dg) = \{(M-s, s)\}$ for $s > 0$, it must be that $\lambda < \lambda_0$. Thus define $\lambda^* = \lambda_0$. For $\mathcal{C}_1(\{dg, e\}) = \{e\}$, we need to establish

$$U(dg) + V(dg) < U(e) + V(e).$$

Observe that

$$U(dg) + V(dg) = 2\left[u(M - s_\lambda, s_\lambda) + \lambda v(M - s_\lambda, s_\lambda)\right] - u\left(\frac{M}{2}, \frac{M}{2}\right) - \lambda v(M, 0)$$

$$\text{and } U(e) + V(e) = u(M, 0) + \lambda v(M, 0)$$

So the desired inequality requires

$$\begin{aligned} &2\left[u(M - s_\lambda, s_\lambda) + \lambda v(M - s_\lambda, s_\lambda)\right] \\ &< u\left(\frac{M}{2}, \frac{M}{2}\right) + u(M, 0) + 2\lambda v(M, 0) \end{aligned}$$

Define the difference

$$\begin{aligned} D(\lambda) &:= U(dg) + V(dg) - [U(e) + V(e)] \\ &= 2u(M - s_\lambda, s_\lambda) - u\left(\frac{M}{2}, \frac{M}{2}\right) - u(M, 0) + 2\lambda\left[v(M - s_\lambda, s_\lambda) - v(M, 0)\right]. \end{aligned}$$

Then $D(0) = u\left(\frac{M}{2}, \frac{M}{2}\right) - u(M, 0) > 0$ and $D(\lambda_0) = u(M, 0) - u\left(\frac{M}{2}, \frac{M}{2}\right) < 0$. Take the first derivative wrt λ, and observe

$$\begin{aligned} \frac{dD(\lambda)}{d\lambda} &= \frac{2du}{ds}\frac{ds_\lambda}{d\lambda} + \frac{2\lambda dv}{ds}\frac{ds_\lambda}{d\lambda} + 2\left[v(M - s_\lambda, s_\lambda) - v(M, 0)\right] \\ &= 2\underbrace{\left(\frac{du}{ds} + \frac{\lambda dv}{ds}\right)}_{=0}\frac{ds_\lambda}{d\lambda} + 2\left[v(M - s_\lambda, s_\lambda) - v(M, 0)\right] \\ &= 2\left[v(M - s_\lambda, s_\lambda) - v(M, 0)\right] < 0 \end{aligned}$$

Therefore, $D(\lambda)$ decreases from $D(0) > 0$ to $D(\lambda_0) < 0$ when λ increases from 0 to λ_0. Conclude that there exists λ_* such that $D(\lambda) < 0$ for $\lambda_* < \lambda < \lambda_0 = \lambda^*$. This completes the proof. ■

Proof of Prop 4. Begin by noting that $U(a) + V(a) = U(\{M - c, 0\}) + V(\{M - c, 0\})$ implies

$$\begin{aligned} &2u(M - s_\lambda, s_\lambda) - u\left(\tfrac{M}{2}, \tfrac{M}{2}\right) - u(M - c, 0) \\ &\qquad + \lambda\left[2v(M - s_\lambda, s_\lambda) - v(M, 0) - v(M - c, 0)\right] = 0 \\ \Longrightarrow\ &u_1(M - c, 0)\, dc + \lambda v_1(M - c, 0)\, dc \\ &\qquad + \left[2v(M - s_\lambda, s_\lambda) - v(M, 0) - v(M - c, 0)\right] d\lambda = 0 \\ \Longrightarrow\ &\frac{dc}{d\lambda} = \frac{v(M, 0) + v(M - c, 0) - 2v(M - s_\lambda, s_\lambda)}{\underbrace{u_1(M - c, 0) + \lambda v_1(M - c, 0)}_{>0}}. \end{aligned}$$

Recall the quantities λ_0 and $D(\lambda)$ defined in the proofs of earlier proposition. Then $s_\lambda = 0$ whenever $\lambda \geq \lambda_0$, and moreover,

$$\begin{aligned} &D(\lambda) > 0 \\ \Longrightarrow\ &U(dg) + V(dg) < U(e) + V(e) \\ \Longrightarrow\ &c > 0. \end{aligned}$$

Thus, $v(M, 0) + v(M - c, 0) - 2v(M - s_\lambda, s_\lambda) = v(M - c, 0) - v(M, 0) < 0$ for $\lambda \geq \lambda_0$, and so $\frac{dc}{d\lambda}\big|_{\lambda \geq \lambda_0} < 0$. By continuity, there exists $\lambda^{**} < \lambda_0$ such that $\frac{dc}{d\lambda} < 0$ for all $\lambda \geq \lambda^{**}$.■

Proof of Prop 7. Given our assumptions on u and v, we have $u(5,5) > u(6,1)$ and $v(6,1) > v(5,5)$. Note that

$U(a_1) + V(a_1) = u(6,5) + \lambda v(6,5)$

$U(a_2)+V(a_2) = \max\{2[u(6,1) + \lambda v(6,1)], 2[u(5,5) + \lambda v(5,5)]\} - [u(5,5) + \lambda v(6,1)]$

$U(a)+V(a) = u(6,l)+\lambda v(6,l) = \frac{1}{2}[u(6,5) + u(6,1)]+\frac{1}{2}\lambda[v(6,5) + v(6,1)]$

Let λ^{***} be the temptation intensity so that in game a_2 the agent is indifferent between choosing the fair choice $(5,5)$ and the unfair choice $(6,1)$. That is, λ^{***} solves $u(6,1)+\lambda^{***}v(6,1) = u(5,5)+\lambda^{***}v(5,5)$. Then $u(6,1)+\lambda v(6,1) > u(5,5) + \lambda v(5,5)$ for any $\lambda > \lambda^{***}$. Therefore, $(6,1)$ is chosen in game a_2 when $\lambda > \lambda^{***}$. Moreover,

$$\begin{aligned}
&U(a) + V(a) - U(\frac{1}{2}, a_1; \frac{1}{2}, a_2) - V(\frac{1}{2}, a_1; \frac{1}{2}, a_2) \\
&= \frac{1}{2}\left\{ \begin{array}{c} u(6,5) + u(6,1) + \lambda v(6,5) + \lambda v(6,1) + u(5,5) \\ -u(6,5) - \lambda v(6,5) - 2u(6,1) - \lambda v(6,1) \end{array} \right\} \\
&= \frac{1}{2}\{u(5,5) - u(6,1)\} \\
&> 0
\end{aligned}$$

Thus $\mathcal{C}_1(\{a, (\frac{1}{2}, a_1; \frac{1}{2}, a_2)\}) = \{a\}$. ■

F Appendix: Proofs of Propositions 5-6

Below we write uv for $u + v$ and similarly for UV. Suppose that:

$$v(l) \leq v(f^*) < v(f_*) \text{ and } u(f_*) < u(l) < u(f^*).$$

Moreover, assume:

(A) $uv(f^*) > uv(l)$ and $uv(f_*) > uv(l)$ (which implies $uv(\alpha f^* + (1-\alpha)f_*) > uv(l)$)

(B) $u(\alpha f^* + (1-\alpha)f_*) < u(l)$

(C) $v(f_*) - v(f^*) \geq 2[u(l) - u(f_*)]$

Some immediate implications are:

$U(\{l, f^*\}) = u(f^*)$, $U(\{l, f_*\}) = u(f_*)$,

$U(\{l, \alpha f^* + (1-\alpha)f_*\}) = u(\alpha f^* + (1-\alpha)f_*)$,

$V(\{l, f^*\}) = v(f^*)$, $V(\{l, f_*\}) = v(f_*) + u(f_*) - u(l)$,

$V(\{l, \alpha f^* + (1-\alpha)f_*\}) = v(\alpha f^* + (1-\alpha)f_*) + u(\alpha f^* + (1-\alpha)f_*) - u(l)$.

Consider the following condition:

$$\textit{Condition } (D): \qquad u(f^*) - u(f_*) \geq v(f_*) - v(f^*)$$

Lemma 9 *(i) For any α in the range where (B) holds, the menu $\{l, \alpha f^* + (1-\alpha)f_*\}$ is the most tempting in $\{\{l\}, \{l, f^*\}, \{l, \alpha f^* + (1-\alpha)f_*\}\}$. In particular, $\{l, f_*\}$ is the most tempting in $\{\{l\}, \{l, f^*\}, \{l, f_*\}\}$.*

(ii) If condition (D) holds (resp. fails), the temptation utility $V(\{l, \alpha f^* + (1-\alpha)f_*\})$ is increasing (resp. decreasing) in α over the range where (B) holds. It is decreasing over the range where it fails.

Proof. Note that $V(\{l, \alpha f^* + (1-\alpha)f_*\}) > V(\{l\})$

iff $v(\alpha f^* + (1-\alpha)f_*) + u(\alpha f^* + (1-\alpha)f_*) - u(l) > v(l)$

iff $v(\alpha f^* + (1-\alpha)f_*) + u(\alpha f^* + (1-\alpha)f_*) > v(l) + u(l)$, which is implied by (A).

Next, we check that

$V(\{l, \alpha f^* + (1-\alpha)f_*\}) \geq V(\{l, f^*\})$

iff $v(\alpha f^* + (1-\alpha)f_*) + u(\alpha f^* + (1-\alpha)f_*) - u(l) \geq v(f^*)$

iff $(1-\alpha)[v(f_*) - v(f^*)] + \alpha[u(f^*) - u(f_*)] \geq u(l) - u(f_*)$ which is implied by (C) and the assumption that $u(f^*) > u(l)$.

This yields claim (i). For (ii), compute that over the range where (B) holds,

$V(\{l, \alpha f^* + (1-\alpha)f_*\})$

$= v(\alpha f^* + (1-\alpha)f_*) + u(\alpha f^* + (1-\alpha)f_*) - u(l)$

$= \alpha v(f^*) + (1-\alpha)v(f_*) + \alpha u(f^*) + (1-\alpha)u(f_*) - u(l)$

$= \alpha[u(f^*) - u(f_*) - (v(f_*) - v(f^*))] + v(f_*) + u(f_*) - u(l)$

The slope wrt α depends on whether the term in the square brackets is positive, which is determined by condition D. This establishes the result. When (B) fails then $V(\{l, \alpha f^* + (1-\alpha)f_*\}) = v(\alpha f^* + (1-\alpha)f_*)$ which is decreasing in α. ■

Lemma 10 $\{\{l\}\} \succ \{\{l\}, \{l, \alpha f^* + (1-\alpha)f_*\}\} \succ \{\{l, f_*\}\}$.

Proof. Since $U(\{l\}) = u(l) > u(\alpha f^* + (1-\alpha)f_*) = U(\{l, \alpha f^* + (1-\alpha)f_*\})$

and $V(\{l, \alpha f^* + (1-\alpha)f_*\}) = v(\alpha f^* + (1-\alpha)f_*) + u(\alpha f^* + (1-\alpha)f_*) - u(l) > v(l) = V(\{l\})$, we therefore have

$U(\{l\}) > W(\{\{l\}, \{l, \alpha f^* + (1-\alpha)f_*\}\}) \geq U(\{l, \alpha f^* + (1-\alpha)f_*\})$. The claim obtains once we observe that

$U(\{l, \alpha f^* + (1-\alpha) f_*\}) = u(\alpha f^* + (1-\alpha) f_*) > u(f_*) = U(\{l, f_*\})$. ■

Consider also:

$$Condition\ (E): \qquad uv(f^*) + [u(l) - u(f_*)] > uv(f_*)$$

Lemma 11 *If condition* (E) *holds then*

$$\{\{l\},\ \{l, f^*\},\ \{l, f_*\}\} \succ \{\{l\}, \{l, \alpha f^* + (1-\alpha) f_*\}\}.$$

If condition (E) *is violated then*

$$\{\{l\}, \{l, \alpha f^* + (1-\alpha) f_*\}\} \succ \{\{l\},\ \{l, f^*\},\ \{l, f_*\}\}.$$

Proof. We have condition (E)
iff $u(f^*) + v(f^*) - [u(f_*) + v(f_*)] > u(f_*) - u(l)$
iff $u(f^*) + v(f^*) - [u(f_*) + v(f_*) - u(l)] > u(f_*)$ and thus
iff $UV(\{l, f^*\}) > UV(\{l, f_*\})$.
Also, $UV(\{l, f^*\}) > UV(\{l\})$ since $uv(f^*) > uv(l)$ is implied by (A). So under (E), $\{l, f^*\}$ is the chosen menu.
If (E) is violated, $\{l, f_*\}$ dominates $\{l, f^*\}$, but in fact it is going to be chosen since
$UV(\{l, f_*\}) \geq UV(\{l\})$
iff $u(f_*) + [v(f_*) + u(f_*) - u(l)] \geq u(l) + v(l)$
iff $v(f_*) - v(f^*) \geq 2[u(l) - u(f_*)]$, which holds by (C).
To summarize, if (E) holds, then
$W(\{\{l\}, \{l, f^*\}, \{l, f_*\}\}) = UV(\{l, f^*\}) - V(\{l, f_*\})$
$> UV(\{l\}) - V(\{l, f_*\}) = u(l) + v(l) - v(f_*) > u(l) = U(\{l\})$.
If (E) is violated then
$W(\{\{l\}, \{l, f^*\}, \{l, f_*\}\}) = UV(\{l, f_*\}) - V(\{l, f_*\}) = U(\{l, f_*\})$.
The conclusion then holds by applying the claim in the previous lemma.
■

Condition (D) implies (E), as (D) implies $[u(f^*) - u(f_*)] + [u(l) - u(f_*)] > v(f_*) - v(f^*)$ which is equivalent to (E).

Lemma 12 *If (D) holds then*

$$\{\{l\},\{l,f^*\},\{l,f_*\}\} \succ \{\{l\},\{l,f^*\},\{l,\alpha f^*+(1-\alpha)f_*\}\}$$

If (E) is violated then

$$\{\{l\},\{l,f^*\},\{l,\alpha f^*+(1-\alpha)f_*\}\} \succ \{\{l\},\{l,f^*\},\{l,f_*\}\}.$$

Proof. If D holds then we must have $UV(\{l,f^*\}) > UV(\{l,\alpha f^*+(1-\alpha)f_*\})$, since

$u(f^*)-u(f_*) \geq v(f_*)-v(f^*)$

$\Longrightarrow (1-\alpha)[u(f^*)-u(f_*)] > (1-\alpha)[v(f_*)-v(f^*)]+u(\alpha f^*+(1-\alpha)f_*)-u(l)$ (by assumption (B))

$\Longrightarrow u(f^*)+v(f^*) > u(\alpha f^*+(1-\alpha)f_*)+v(\alpha f^*+(1-\alpha)f_*)+u(\alpha f^*+(1-\alpha)f_*)-u(l)$

$\Longrightarrow UV(\{l,f^*\}) > UV(\{l,\alpha f^*+(1-\alpha)f_*\})$.

D implies (E), and so $UV(\{l,f^*\}) > UV(\{l\})$ also holds as in the proof of a previous lemma. Therefore,

$W(\{\{l\},\{l,f^*\},\{l,\alpha f^*+(1-\alpha)f_*\}\}) = UV(\{l,f^*\})-V(\{l,\alpha f^*+(1-\alpha)f_*\})$,

and we saw in a previous lemma that (E) also implies $W(\{\{l\},\{l,f^*\},\{l,f_*\}\}) = UV(\{l,f^*\})-V(\{l,f_*\})$.

By an earlier lemma, for α where (B) holds, condition (D) implies $V(\{l,\alpha f^*+(1-\alpha)f_*\}) > V(\{l,f_*\})$, and therefore:

$$W(\{\{l\},\{l,f^*\},\{l,f_*\}\}) > W(\{\{l\},\{l,f^*\},\{l,\alpha f^*+(1-\alpha)f_*\}\}),$$

as desired.

If E is violated then

$W(\{\{l\},\{l,f^*\},\{l,f_*\}\})$

$= U(\{l,f_*\})$ (as shown in the proof of earlier lemma)

$= u(f_*)$

$< u(\alpha f^*+(1-\alpha)f_*)$

$= U(\{l,\alpha f^*+(1-\alpha)f_*\})$

$\leq W(\{\{l\},\{l,f^*\},\{l,\alpha f^*+(1-\alpha)f_*\}\})$ (since $\{l,\alpha f^*+(1-\alpha)f_*\}$ is the most tempting by earlier lemma).

This completes the proof. ■

For the example in the text it is straightforward to check that condition (D) is satisfied iff $m \geq 1$, and condition (E) is satisfied iff $m > 0.75$.

References

[1] Alaoui, L. (2012): "The Value of Useless Information", mimeo.

[2] Aliprantis, C. and K. Border (1994): *Infinite Dimensional Analysis: a Hitchhiker's Guide, 2nd Edition*, Springer Verlag.

[3] Andreoni, J., and D. Bernheim (2009), 'Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects', Econometrica 77(5), pp. 1607-36.

[4] Andreoni, J., J. Rao and H. Trachtman (2012): "Avoiding The Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving," NBER working paper 17648.

[5] Batson, C., D. Kobrynowicz, J. Dinnerstein, H. Kampf and A. Wilson (1997): 'In a Very Different Voice: Unmasking Moral Hypocrisy', *Journal of Personality and Social Psychology* 72, pp. 1335-1348.

[6] Battigalli and Dufwenberg (2007): 'Guilt in Games', *American Economic Review* 97(2), pp 170-176.

[7] Bandura, A (1977): *Social learning theory*, Englewood Cliffs, NJ: Prentice Hall.

[8] — (1991): 'Social Cognitive Theory of Moral Thought and Action', in W. Kurtines and J. Gewirtz (eds), *Handbook of Moral Behavior and Development, volume 1: Theory*, pp 45-103. Hillsdale, NJ: Erlbaum.

[9] Border, K. (1985): 'More on Harsanyi's Utilitarian Cardinal Welfare Theorem', *Social Choice and Welfare* 1, pp 279-281.

[10] Breman, A. (2011): 'Give More Tomorrow: Two Field Experiments on Altruism and Intertemporal Choice', Journal of Public Economics 95 (11–12), pp 1349–1357.

[11] Broberg, T., Ellingsen, T., and Johannesson, M. (2007), 'Is Generosity Involuntary?', *Economics Letters* 94(1), pp. 32-37.

[12] Dana, J., D. Cain and R. Dawes (2006): 'What You Don't Know Won't Hurt Me: Costly (But Quiet) Exit in Dictator Games', *Organizational Behavior and Human Decision Processes* 100, pp 193-201.

[13] Dana, J., R. Weber and J. Kuang (2007): 'Exploiting Moral Wiggle Room: Experiments Demonstrating and Illusory Preference for Fairness', *Economic Theory* 33, 67-80.

[14] Della Vigna, S., J. List, and U. Malmendier (2012), 'Testing for Altruism and Social Pressure in Charitable Giving,' *Quarterly Journal of Economics* 127, pp. 1-56.

[15] Della Vigna, S., and U. Malmendier (2006), 'Paying Not to Go to the Gym,' *American Economic Review* 96(3), pp 694-719.

[16] Dekel, E., B. Lipman and A. Rustichini (2001): 'Representing Preferences with a Unique Subjective State Space', *Econometrica* 69, pp 891-934.

[17] Dekel, E., B. Lipman and A. Rustichini (2005): 'Temptation-Driven Preferences', *Review of Economic Studies* .

[18] Dillenberger and Sadowski (2012): 'Ashamed to be Selfish', *Theoretical Economics* 7(1), pp 99-124.

[19] Ehrich, K., and J. Irwin (2005): 'Willful Ignorance in the Request for Product Attribute Information', *Journal of Marketing Research* 42(3), pp. 266-277.

[20] Ellingsen, T., M. Johannesson, S. Tjøtta, and G., Torsvik (2010), 'Testing Guilty Aversion', *Games and Economic Behavior* 68(1), pp. 95-107.

[21] Enns, M., B. Cox, J. Sareen and P. Freedman (2001): "Adaptive and Maladaptive Perfectionism in Medical Students: A Longitudinal Approach," *Medical Education* 35, pp1034-1042.

[22] Ferrari, J., J. Johnson and W. McGown (1995): *Procrastination and Task-Avoidance: Theory, Research and Treatment*. New York: Plenum Press.

[23] Fudenberg, D., and D. Levine (2006): 'A Dual Self Model of Impulse Control', *American Economic Review* 96, pp. 1449-1476.

[24] Fudenberg, D., and D. Levine (2010): 'Risk, Delay, and Convex Self-Control Costs', mimeo.

[25] Fudenberg, D., and D. Levine (2012): 'Timing and Self-Control', *Econometrica* 80(1), pp 1-42.

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

[27] Gul, F. and W. Pesendorfer (2003): 'Self-Control and the Theory of Consumption', *Econometrica* 72, pp 119-158.

[28] Hammam, J., G. Loewenstein and R. Weber (2009): 'Self-Interest Through Delegation: An Additional Rationale for the Principal-Agent Relationship', mimeo.

[29] Kopylov, I. (2009): 'Temptations in General Settings', *B.E. Journal of Theoretical Economics* 9(1), article 31.

[30] Kopylov, I. (2012): 'Perfectionism and Choice', *Econometrica* 80(5), pp 1819-1843.

[31] Kopylov, I. and J. Noor (2009): 'Self-Deception and Choice', mimeo.

[32] Kreps, D. and E. Porteus (1978): 'Temporal Resolution of Uncertainty and Dynamic Choice Theory' *Econometrica* 46(1), pp. 185–200.

[33] Laibson, D. (1997): "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics* 112, pp 443-77.

[34] Lazear, E., U. Malmendier and R. Weber (2006): 'Sorting in Experiments with Application to Social Preferences', *American Economic Journal: Applied Economics* 4(1), pp. 136-63.

[35] Neilson, W. (2009): 'A Theory of Kindness, Reluctance, and Shame for Social Preferences', *Games and Economic Behavior* 66(1), pp 394-403.

[36] Noor, J. (2007): 'Commitment and Self-Control', *Journal of Economic Theory* 135, pp 1-34.

[37] Noor, J. (2011): 'Temptation and Revealed Preference', *Econometrica* 79(2), pp601-644

[38] Noor, J. and N. Takeoka (2009): 'Menu-Dependent Self-Control', mimeo.

[39] O'Donoghue, T., and M. Rabin (1999), 'Doing it Now or Later', *American Economic Review* 89, pp 103-24.

[40] Saito, K. (2015): 'Impure Altruism and Impure Selfishness', forthcoming in *Journal of Economic Theory*.

[41] Sarver, T. (2008): 'Anticipating Regret: Why Fewer Options may be Better', *Econometrica* 76(2), pp 263–305.

[42] Strotz, R. (1955), "Myopia and Inconsistency in Dynamic Utility Maximization," *Review of Economic Studies* 23, pp 165-180.