Temptation and Guilt*

Jawwad Noor Linxia Ren

January 9, 2011

Abstract

This paper builds a model of guilt on the observation that guilt-avoiding behavior must itself be cause for guilt. The agent experiences guilt when she submits to any temptation, but the anticipation of guilt in turn generates a temptation to avoid guilt. In an application to social preferences, where the literature explains behavior by a concern for social image, the model uses temptation and guilt by selfishness to accommodate all key findings in experiments on dictator games. The model also yields welfare implications different from those in the literature. Results are reported from an experiment that suggests the existence of a temptation to be selfish.

Keywords: Guilt, normative preference, temptation, social preference, dictator games, welfare.

1 Introduction

Guilt is the painful experience that accompanies a ‘bad’ choice. This definition presumes the existence of a normative preference – the agent’s personal

*Noor (email: jnoor@bu.edu) and Ren (email: lxren@bu.edu) are at the Dept of Economics, Boston University, 270 Bay State Road, Boston MA 02215. We thank Bart Lipman, Philipp Sadowski, Mike Luca, Rob Munger, Sean Horan, seminar participants at Boston University 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.
view of what should be chosen – which is the standard against which a choice is judged to be good or bad. The definition also presumes the existence of factors that may cause choice to deviate from normative preference. These factors may be attributed to the agent’s desires – refer to these as her temptation preferences. The agent’s choices are thus a struggle and a search for a compromise between normative and temptation preferences. When normative and temptation preferences conflict, pain is inevitable: good choices require the painful exertion of self-control, but bad choices entail the painful experience of guilt.

1.1 A Theory of Guilt

What behavior reveals that an agent experiences guilt? A response that comes naturally to economists is that the experience of any painful emotion gives rise to pain-avoiding behavior. Thus, it is natural to hypothesize that the behavioral implication of guilt is guilt-avoidance. Evidence of guilt-avoidance is suggested by experiments in social psychology and in economics. Ehrich and Irwin [16] 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.

While it may seem natural that a model of guilt should be built around guilt-avoidance just as a model of any painful emotion should be built around pain-avoidance, this paper takes the view that guilt-avoidance is a different phenomenon compared to usual forms of pain-avoidance. An agent’s action to avoid (say) risk can be taken at face value: the fact that risk is painful to bear is a complete explanation for risk-avoiding behavior and has the obvious welfare implications. We hold that this does not necessarily hold true for guilt-avoidance. While pain-avoidance is typically never itself painful, we argue that guilt-avoidance should itself be cause for guilt. If guilt is caused by a betrayal of one’s resolve toward normative goal, then guilt-avoidance is such a betrayal as well. In the context of the earlier example, if an agent considers it wrong to knowingly consume an unethical product, he must also consider it wrong to willfully avoid information on the ethicality of a product.
In a word, while it is natural (and also suggested by experimental evidence, such as Ehrich and Irwin [16]) that agents who experience guilt also desire to avoid guilt, we believe that a complete theory of guilt should also explain why guilt-avoidance might itself generate guilt. In fact we believe it is necessary that it does. Other than completing the conceptual framework, the answer will suggest both the behavioral implications and welfare implications of the desired theory.

We make sense of ‘guilt by guilt-avoidance’ in the following way: We hypothesize that the desire to avoid guilt is not a normative goal, but rather itself a temptation. We hypothesize that it arises because opportunities of guiltless indulgence are more tempting than those of guilt-ridden indulgence. This induces a temptation to avoid guilt. It also completes the picture: It explains why a moral agent may try to avoid guilt, and it explains why guilt-avoiding behavior may itself give rise to guilt. (We discuss implications for behavior and welfare below in the context of our application.)

Based on this analysis, we model an agent whose choices are determined by the maximization of an ‘aggregation’ of normative and temptation preferences less guilt costs and self-control costs. This describes both how an agent chooses from a menu (choice problem) and also how he selects a menu for himself to face in the future. The normative preference over menus takes into account the normative value of the choice in a menu and the associated self-control costs – in particular, guilt-avoidance is not a motive. The temptation preference over menus takes into account the temptation value of the choice in a menu and the associated guilt costs – guilt-avoidance is a motive here.\(^1\)

This paper contributes to the literature on self-control problems (Laibson [29], Gul and Pesendorfer [22]) by studying what we consider to be an integral part of the problem of dealing with temptation, and by identifying a conceptual nonlinearity associated with it.\(^2\) Our second contribution is the

---

\(^1\)The formal framework for studying temptation introduced by Gul and Pesendorfer [22] is now sufficiently well-developed in the decision theory literature to accommodate our theory of guilt (Noor [34], Kopylov [26], Kopylov and Noor [28]). We outline our formal model in Section 2 and provide behavioral foundations in Section 5.

\(^2\)This paper has limited overlap with the work on guilt by Battigalli and Dufwenberg [5]. 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, which could in principle depend on expectations, though we do not model this. (See Ellingsen et al [17] for experimental results that suggest insignificant affect of others’ expectations on behavior.) Another
application detailed below. While much theoretical work has been done on models of agents with the propensity for self-control (see Gul and Pesendorfer [22] and subsequent literature, Fudenberg and Levine [19]) there have been relatively few demonstrations of the empirical relevance of the propensity of self-control (Fudenberg and Levine [19, 20, 21], Noor and Takeoka [35]). This is in contrast with the many applications of the dynamic inconsistency associated with temptation. We provide a demonstration of the relevance of the notion of self-control in the context of social preferences.

1.2 Application to Social Preferences

We apply our model to the literature on social preference. The application exploits the behavior generated by the model, which includes guilt-avoidance. ‘Guilt by guilt-avoidance’ remains in the background, manifesting itself only in our stance on welfare.

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 [6, 7]). 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 [4]). Some of the evidence is as follows:

- In seminal work in psychology, Batson et al [4] 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

\[\text{difference is that Battigalli and Dufwenberg [5] focus on strategic settings whereas we focus on a single decision maker.}\]
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 [30] 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 [10] find that a significant proportion exit the game even at a cost. Hammam et al [24] 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 [11] 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.

While initial research on fairness and social preferences suggested the existence of an intrinsic concern for altruism and fairness, the newer research undermines this conclusion. It suggests that observed moral behavior is not indicative of the true extent of an agent’s moral imperative, but rather overstates it because it may be the result of social pressure. This view has gained some favor in economics, particularly in the study of social preferences and the dictator game (see also Neilson [32], Andreoni and Bernheim [2], and Della Vigna et al [12]). Our model of guilt suggests a different relationship between observed behavior and unobserved moral imperative: Choices reflect the interaction of a personal moral imperative and the temptation to be selfish. Behavior in the absence of social pressure understates the extent of an agent’s moral imperative because of the heightened influence of temptation. Social pressure may enable an agent to better align his choices with his moral perspective. We show that the key comparative statics observed in experiments on the dictator game are all natural implications of our model.

Does it matter whether the world is viewed as populated by amoral moral hypocrites rather than by temptation-stricken moralists? The answer is yes, on both positive and normative grounds. (The following discussion also highlights the peculiar behavioral and welfare implications of our theory of guilt).
On the positive side, we show that the moral hypocrisy and moral integrity are not observationally equivalent. Specifically, we argue that to the extent that choices from a ‘distance’ (by some appropriate measure of distance) are less subject to temptation, temptation-stricken moralists would behave more morally from a distance. In particular, their behavior may change with distance. In contrast the behavior of moral hypocrites should remain unchanged, as the driving force behind their behavior is how they appear to others, which is not a function of distance. We conduct a version of the dictator game experiment and find that behavior indeed becomes more generous with temporal distance.

On the normative side, there is the question of whether excuses or opportunities for immoral behavior should be provided. The welfare implication of moral hypocrisy is that agents may be better off if they can avoid the costs of behaving morally. Thus, Batson et al’s subjects are better off when they are given the coin, and dictators in dictator games are better off when they are given the silent exit option. In contrast, temptation-stricken moralists are worse off, since the coin or the silent exit option puts them in the presence of temptation. Indeed, social pressure may make the agent better-off, as it may serve as a commitment device.

The remainder of the paper is organized as follows. Section 2 presents our model of guilt. Section 3 discusses the application to dictator games. Section 4 presents the results of our experiment. Section 5 provides behavioral foundations for our model and also defines and characterizes a comparative measure of proneness to guilt. Section 6 discusses related decision-theoretic literature. Section 7 concludes. All proofs are contained in appendices.

2 A Model of Guilt

Denote the space of alternatives by $\Delta$. A menu (choice problem) is a non-empty subset of $\Delta$, and the set of menus is denoted by $\mathcal{M}$. Generic elements of $\Delta$ are $x, y, z$ while those of $\mathcal{M}$ are $a, b, c$.

The agent has a personal view of how she should behave when faced with a menu. This view may reflect ethical concerns, but more generally may be her perspective on what is the right balance between social norms and self-interest, for instance. We refer to this perspective on how she should behave as her normative preference. Her actual choices are also influenced
by desires – refer to these as her temptation preferences. Choices are thus a struggle and a search for a compromise between normative and temptation preferences. When normative and temptation preferences conflict, pain is inevitable: good choices require the painful exertion of self-control, but bad choices are accompanied by the painful experience of guilt.

Observe that the notions of normative and temptation preferences are necessary for any discussion of guilt. Since foundations for the notions of normative and temptation preferences have been provided in the work of Gul and Pesendorfer [22] (henceforth GP), we take inspiration from it and write down an extension of their model that accommodates guilt.

2.1 Choice

We provide here an outline of the model and specify how choice is determined by it. For the theoretically-inclined readers: the formal model and its behavioral foundations are presented in Section 5.

Choice from a Menu. Normative and temptation preferences over alternatives are represented by utilities $u, v : \Delta \rightarrow \mathbb{R}$ respectively. When faced with a menu $a \in \mathcal{M}$, the agent’s choice $C(a)$ from the menu is the set of maximizers:

$$\arg\max_{x \in a} \left\{ u(x) + v(x) - \left[ \max_{z \in a} u(z) - u(x) \right] - \left[ \max_{y \in a} v(y) - v(x) \right] \right\} \quad (1)$$

$$= \arg\max_{x \in a} \left\{ u(x) + v(x) \right\}. \quad (2)$$

We first interpret (1) and then verify that it equals (2). Begin with some initial observations: The normatively best choice in the menu $a$ solves $\max_{z \in a} u(z)$ whereas the most tempting choice solves $\max_{y \in a} v(y)$. An alternative $x \in a$ will satisfy normative and temptation preferences to varying degrees. To the extent that $x$ may not respect normative preferences, the agent suffers the cost of guilt $\left[ \max_{z \in a} u(z) - u(x) \right] \geq 0$. If $x$ does not satisfy desires, she feels the frustration measured by $\left[ \max_{y \in a} v(y) - v(x) \right] \geq 0$. This can be interpreted as the cost of self-control, which by definition must be proportional to the degree of frustration of desires.

Given these observations, (1) is interpreted as follows: a choice $x$ from menu $a$ yields normative utility $u(x)$ less guilt costs, and temptation utility $v(x)$ less self-control costs, and the agent maximizes the sum of these costs.
Observe that actual costs are not of any consequence in describing choice from a menu. Leaving aside what she ‘experiences’ in the menu, the agent’s choice at the end of the day simply maximizes the sum of normative and temptation utilites: since \( \max_{z \in A} u(z) \) and \( \max_{z \in A} v(z) \) are constants when \( a \) is given, they are sunk at the time of choice and thus do not determine choice. Thus, her choice from a menu seeks a compromise between her normative perspective and her desires – this is reflected in (2).\(^3\)

The experience of guilt and self-control costs reveals itself in a different domain, to which we now turn.

**Choice between Menus.** Our study will involve not only a discussion of how an agent would choose from menus but also how she would choose between menus. This will be required, for instance, to model the ‘exit option’ in experiments on the dictator games (see the Introduction), and also to capture the affect of guilt and self-control costs. Below we will refer to choice of menu as taking place in some period 1, and subsequent choice from a menu as taking place in period 2.

\[
\begin{align*}
&\text{t=1} \quad \text{choose menu } a \\
&\text{temptation to avoid ex post guilt} \\
&\text{t=2} \quad \text{choose alternative } x \in a \\
&\text{temptation and guilt}
\end{align*}
\]

The kind of considerations entering period 2 choice will exist also in period 1. Thus, the agent will be presumed to have a normative preference and temptation preference over menus, just as she does over alternatives as above, and her choice of menu will be subject to temptation and guilt as well. This is formalized as follows: given a menu of menus \( A \), her choice is

\[
\begin{align*}
C_1(A) &= \arg \max_{a \in A} \left\{ U(a) + V(a) - \left[ \max_{b \in A} U(b) - U(a) \right] - \left[ \max_{c \in A} V(c) - V(a) \right] \right\} \\
&= \arg \max_{x \in A} \left\{ U(a) + V(a) \right\}.
\end{align*}
\]

That is, her choice of menu seeks a compromise between her normative and temptation preferences over menus, which are represented by \( U \) and \( V \) resp, and described next.

\(^{3}\)Though still giving rise to (2), an alternative to (1) that is familiar from the literature on temptation is considered at the end of Section 6.
Normative and temptation preferences over menus will be derivative concepts. Her normative preference over menus is represented by a utility $U : \mathcal{M} \rightarrow \mathbb{R}$ of the GP form

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

The agent cares about three things: her time 2 choice from that menu, the normative utility of that choice, and the self-control cost associated with it.\(^4\) Thus, she normatively prefers menus that have normatively better choices and lower self-control costs. A key hypothesis here is that *she does not care to avoid guilt costs*. We justify this hypothesis in next Section.

Her temptation preference over menus is represented by a utility $V : \mathcal{M} \rightarrow \mathbb{R}$ of the form

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

The parameter $\kappa > 0$ represents her relative time discounting (we normalize the normative discount factor to 1). Like the normative perspective, the temptation value of a menu is sensitive to three things: her period 2 choice from that menu, the temptation utility of that choice, and the guilt cost associated with it.\(^5\) Thus, she is tempted by menus that lead to more tempting choices and lower guilt. In particular, this expresses our key hypothesis that *guiltless indulgence is more tempting than guilt-ridden indulgence*. Intuitively, guilt dulls the pleasure from the consumption of tempting alternatives. For instance, it may be harder to enjoy an expensive dinner when a beggar (an option to be charitable) is visible from the window. Hence, in period 1 the agent is tempted to avoid guilt.

\(^4\)It is worth observing the internal consistency of the model. Note that

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

Therefore, the maximizer $x \in a$ in the definition of $U$ is indeed the period 2 choice from $a$.

\(^5\)Note again the the internal consistency of the model:

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

Therefore, the maximizer $x \in a$ in the definition of $V$ is indeed the period 2 choice from $a$.\(^{9}\)
This completes our description of the model. Observe that, analogous to period 2, the choice (3) is such that self-control and guilt costs associated with the choice of menu do not express themselves in period 1, except through the implication that the agent seeks a compromise. Note however, that period 2 self-control and guilt costs express themselves in period 1 choice, since they enter $U$ and $V$. The agent’s normative preference underlying $U$ strives to avoid temptation and self-control costs while the temptation preference underlying $V$ seeks temptation but avoids guilt costs.

2.2 Welfare and Guilt-Avoidance

In our model, the agent’s welfare is defined by her normative perspective. That is, her normative perspective is her personal welfare criterion, and it indicates what she feels the ‘best’ thing to do. Her normative perspective is $u$ in period 2 and $U$ in period 1. Unlike traditional revealed preference theory, there is a wedge between choice and welfare – choice in any period is ‘contaminated’ with temptation. Unlike behavioral economics, the agent’s “experienced utility” is not her standard for welfare. For instance, guilt costs are not a consideration for our agent’s normative preference. This last claim requires justification, to which we now turn.

For concreteness, suppose that there is a ‘good’ option $g$ and a ‘bad’ option $b$. Suppose further that the agent submits to the temptation by $b$ in period 2 when faced with a choice from $\{g, b\}$:

$$C(\{g, b\}) = \{b\}.$$  

In period 1, she anticipates her choice and recognizes that her choice will be accompanied by pangs of guilt. Contrary to our hypothesis, suppose that her normative perspective takes the view that she should avoid guilt. Then the key observation to make is that in period 1, for some option $B$ that is ‘worse’ than $b$, it will be the case that $\{B\}$ is normatively preferred to $\{g, b\}$. The reason is that even if $B$ is a worse option, its consumption is guiltless, whereas the bad option is chosen with guilt in $\{g, b\}$. Therefore, if a guilt-avoidance motive exists in normative preference, then it would make it normatively acceptable for to commit to a worse option. We find this

\footnote{Observe that in a singleton menu $\{B\}$, the only option is to consume $B$ therefore there is no room to feel guilty. Guilt would be experienced if a better option was a choice, as it was in $\{g, b\}$.}
uncompelling as a feature of normative preference. If guilt is caused by a betrayal of one’s resolve towards normative goals, then guilt-avoidance – such as in this example – involves precisely such a betrayal. If the agent believes she should be moral/other-regarding/fair, then the act of committing to being immoral/selfish/unfair is itself immoral/selfish/unfair. In a moral context, if being immoral is cause for guilt, then it is also true that ‘to want to be immoral is to be immoral’ (Elster [18, pg 65]). Indeed, we believe that guilt-avoidance should itself be cause for guilt, and thus that it cannot be a feature of normative preference.

In our model, a guilt-avoidance motive does indeed exist, but its source is temptation preferences, not normative preferences.

3 Application to Dictator Games

In this section we show that an appropriately specialized version of our model can 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 \(\lambda v\), where \(\lambda\) is a scalar.

\[\text{An illustration in an interpersonal context also helps make the point: guilt-avoidance is like a parent forcing his child to smoke so that the child does not feel guilty about being a smoker.}\]
that parametrizes the intensity of temptation. Note that the dictator’s choice from any menu can be written as

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

\[ = \arg \max_{x \in a} \{ u(x) + \lambda v(x) \}. \]

Evidently, greater intensity \( \lambda \) of temptation is associated with lower self-control. Finally, the dictator’s choice over menus is given by

\[ C_1(A) = \arg \max_{x \in a} \{ U(a) + V(a) \}. \]

where the parameter \( \kappa \) 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 1** For a dictator with temptation intensity \( \lambda \), denote the choice by

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

Then \( s_\lambda \) is decreasing in \( \lambda \).

### 3.1 Evidence Accommodated by Model

We summarize evidence and show that our model can accommodate it.

**Dictators tend to share though they share less than 50%**: In our model, since \( C(a) \) 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 [10] find that over 25% of dictators exit at a cost (specifically, they exit with $9 when the pie is worth $10). Broberg et al [9] find that 64% of their dictators exhibit willingness to exit the game for as little as 82% of the pie.
In our model, since \( C_1(A) \) 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 2** There exists \( \lambda^*, \lambda_* \) s.t. for all \( \lambda_* \leq \lambda \leq \lambda^* \),

\[
C(dg) = \{(M - s, s)\}, \quad s > 0, \text{ and } C_1(\{dg, e\}) = \{e\}
\]

The proposition tells us that there are values of \( \lambda \) 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 [10] 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 [9] find that subjects who offer nothing also value exit less than those subjects who offer positive amounts.

In our model if \( C(a) \) offers nothing then that is indicative of high \( \lambda \). That is, the intensity of a temptation to be selfish is high. For the same agents, \( C_1(A) \) would care less for commitment: observe that \( V \) maximizes temptation utility less guilt costs and that higher \( \lambda \) 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 3** There exists \( \lambda^{**} \) s.t. \( \frac{dc}{d\lambda} < 0 \) 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 $\lambda^{**}$ is strictly less that the $\lambda^*$ in Proposition 2. Agents with temptation intensity $\lambda^{**} \leq \lambda \leq \lambda^*$ share with recipients, but the more they share, the higher their temptation preference for exiting.

**Dictators prefer less information:** Dana et al [11] show that subjects behave more selfishly when they have ‘moral wiggle room’. In their experiment, subjects had the option of playing a random version of the dictator game:

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

where $\ell = (\frac{1}{2}, 5; \frac{1}{2}, 1)$ 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 as follows. 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 $\lambda$ is high (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 overwhelming. Denote by $(\frac{1}{2}, a_1; \frac{1}{2}, a_2)$ then lottery over $a_1$ and $a_2$.

**Proposition 4** There exists $\lambda^{***}$ s.t. $C_1(\{a, (\frac{1}{2}, a_1; \frac{1}{2}, a_2)\}) = \{a\}$ for all $\lambda \geq \lambda^{***}$.

### 3.2 Accommodating the Effect of Anonymity

In order to accommodate the finding that dictators’ choices are different under anonymity, we consider an enriched version of our model by hypothesizing that $\lambda$ is in fact a function:

$$\lambda = \lambda(o),$$
where $o$ is some measure of how observable the agent’s choice is to the recipient. We hypothesize that intensity of temptation is decreasing in observability.\footnote{An axiomatization of this extended model would involve that the ex ante preference $\succeq_o$ be indexed by observability $o$. Then a condition across $\succeq_o$’s is required that reflects the idea that self-control reduces with $o$.}

**Dictators share less under anonymity:** Dana et al [10] find that while only 24% of dictators offered nothing to receivers in the standard dictator game with no anonymity, 46% offered nothing under anonymity. Similarly, Lazear et al [30] find that only 19% of dictators offered nothing under no anonymity, but 33% offered nothing under anonymity. For a related experiment that varies ‘social distance’, see Hoffman et al [25].

In our model, since $C(a)$ maximizes $u + \lambda v$ and since $\lambda$ is hypothesized to increase under anonymity, it follows that choices become more selfish with anonymity.

**Dictators seek anonymity:** In Andreoni and Bernheim [2], the players were given the endowment $(20, 0)$. The dictator had to decide how much to share. His choice was enforced with probability $1 - p$, and with remaining probability nature would determine the allocation. Specifically, nature would enforce each of $(19, 1)$ or $(1, 19)$ with probability $\frac{p}{2}$. Recipients were not informed of the share offered by the dictator. It was found that the fraction of subjects offering $(19, 1)$ increased with $p$ and those offering more fair splits decreased.

Two observations are relevant here to see that our model can accommodate this. First, while unfair allocations yield guilt, as $p$ increases the expected guilt reduces. Thus, dictators would be more likely to make unfair allocations. Second, the observability $o$ of the agent’s choices is a function of the share offered. For instance, if the agent offers $(20, 0)$ and this is enforced, then recipients will deduce the agent’s offer, but if the agent offers $(19, 1)$ then the recipient is less sure whether this was due to a move by nature or because it was the dictator’s offer. Since $\lambda$ increases with lower $o$, the agent is tempted to lower $o$. Indeed, as $p$ reduces, attempts to maximize temptation strengthen (due to lower expected guilt cost) and this includes attempts at lowering $o$. 

3.3 Accommodating Menu-Dependence

In order to accommodate evidence of menu-dependence of dictators’ choices, we consider an enriched version of our model by hypothesizing that $\lambda$ is a function of the form

$$\lambda = \lambda(\max_{y \in a} v(y))$$

where $\max_{y \in a} v(y)$ is the temptation value of the most tempting option available in the menu $a$. Observe that this introduces menu-dependence into $\lambda$. We hypothesize that intensity of temptation is increasing in $\max_{y \in a} v(y)$. This says that self-control becomes weaker in the presence of greater temptation.

**Dictators give less when taking is also an option:** List [31] and Bard-sley [3] find that giving is menu-dependent. For instance, List [31] first considered a standard dictator game where the endowment was ($10,$5) and dictators could share up to $5 of their endowment with the recipient:

$$a_1 = \{(10 - s, 5 + s) : 0 \leq s \leq 5\}.$$  

The mean offer among dictators was $1. However, when given also the option of taking exactly $1 from the recipients,

$$a_2 = a_1 \cup \{(11, 4)\},$$

few dictators took the new option but the rate of giving substantially declined, and the mean offer fell to $0.

Our model would explain this as follows: dictators are tempted to be selfish, but this temptation becomes more intense when more selfish opportunities are offered. Thus, the introduction of the taking option is accompanied by a weakened resolve toward normative goals, which induces the menu-dependence of choice.

Formally, observe that $\max_{x \in a_1} v(x) = v(10, 5) < v(11, 4) = \max_{x \in a_2} v(x)$. Thus $\lambda$ is higher at $a_2$ than at $a_1$, and consequently the choice from $a_2$ is more selfish than that from $a_1$.

---

9 An axiomatization of this extended model would weaken Independence as in Noor and Takeoka [35] to accommodate such menu-dependent self-control.

10 More precisely, they could share only in $0.5$ increments. We abstract away from this in the description.
4 Temptation and Social Preferences: An Experiment

We have shown that our model can accommodate findings in social preference experiments. Each of the findings have alternative explanations in terms of a concern for social image (see the cited references). For instance, such a concern can clearly rationalize the finding that subjects give when playing the dictator game but also prefer a silent exit. In this section we devise a simple experiment that separates the social image theory from our temptation theory of social preference. To be clear, our goal is to show that temptation is relevant for social preference theory, not that social image is irrelevant.

We seek an instrument which responds to temptation but is otherwise unrelated to social image concerns. An idea prevalent in the temptation literature is that immediate temptation is harder to resist than distant temptation (Noor [34]). Put differently, choices from a distance more closely reveal normative preferences. Thus, changes in behavior (“reversals”) that arise due to distancing are attributable to temptation (depending on the context, of course). On the other hand, a concern for social image will not induce a change in behavior with distance as long as the degree of visibility of choices to others is held constant. We use these ideas in the following way.

Design. Our experiment consists of two dictator game treatments. In each treatment, dictators must divide $10 between themselves and the American Red Cross. Treatment A is a standard dictator game. The payments to both parties are made immediately at the end of the experiment – the dictators are paid in cash and the charity receives payment electronically. In treatment B, these payments are made one month later.

Dictators’ responses were anonymous (their sealed cash payments were connected to their identities by a code). Thus, we tried to minimize the impact of social image concerns. More importantly, however, we were careful to equalize any residual impact across the two treatments. This was done by ensuring that the protocols used in each treatment were identical in every respect except for the difference in payment dates.

The experiment was conducted with undergraduates at Boston University that were taking summer economics courses. There were 22 subjects in treatment A and 26 in treatment B.

Results. Consistent with what a temptation theory of social preference
would predict, we found that dictators on average gave less in treatment A than in treatment B. That is, dictators behaved significantly more selfishly when immediate cash was at stake. From a (temporal) distance, dictators were substantially more generous, with the average contribution almost double that without distance: the average donation out of $10 by dictators was $3.875 (std 4.28) in treatment A and $6.529 (std 3.81) in treatment A. This suggests the existence of a temptation to be selfish and the existence of a substantial normative preference to give. Observe that while the average giving rate was less than 50% in treatment A (which is consistent with what is typically observed in dictator game experiments), the average giving rate was more than 50% in treatment B.

<table>
<thead>
<tr>
<th>Dictators’ payment $x$</th>
<th>Group A (%)</th>
<th>Group B (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0$</td>
<td>18.2</td>
<td>42.3</td>
</tr>
<tr>
<td>$0 &lt; x &lt; $5$</td>
<td>13.6</td>
<td>11.5</td>
</tr>
<tr>
<td>$x = $5$</td>
<td>9.1</td>
<td>23.1</td>
</tr>
<tr>
<td>$5 &lt; x &lt; $10$</td>
<td>22.7</td>
<td>9.1</td>
</tr>
<tr>
<td>$x = $10$</td>
<td>36.4</td>
<td>15.4</td>
</tr>
</tbody>
</table>

Formally, we reject the null hypothesis that the distribution of dictators’ behavior was independent of distance. We performed three tests for the differences in distributions of the two samples: the Kolmogorov-Smirnov test, Wilcoxon rank-sum (Mann-Whitney) test and Epps-Singleton test. The test results are summarized in Table 1. All three test rejected the null hypothesis at 10% level, with the smallest p value being 0.025 (rank-sum) and the largest being 0.063 (Kolmogorov-Smirnov).

<table>
<thead>
<tr>
<th>Tests</th>
<th>Test Statistic</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kolmogorov-Smirnov test</td>
<td>0.3601</td>
<td>0.063</td>
</tr>
<tr>
<td>Wilcoxon rank-sum test</td>
<td>-2.241</td>
<td>0.0250</td>
</tr>
<tr>
<td>Epps-Singleton test</td>
<td>10.669</td>
<td>0.03055</td>
</tr>
</tbody>
</table>

A noteworthy observation is that there are factors that may push subjects to be less generous in treatment B, and thus our results potentially
underestimate the extent of subjects’ generosity from a distance. Consider the following:

(1) Charity payments are more valuable when made immediately than later. Therefore, there is an incentive for concerned subjects in treatment B to pay charity immediately out of their own pockets today, and simply demand $10 after a month.

(2) Subjects in treatment B could have saved on cognitive costs or retained flexibility in the face of uncertainty by opting to receive $10 after a month, and then at that time deciding how to allocate it between themselves and charity.

A possible alternative explanation for our finding is the existence of a warm glow. In either treatment, the warm glow from giving may be immediate, in which case it would be relatively more important than delayed personal consumption in treatment B, thus leading to higher contribution rates. A simple experimental test of this possibility is to conduct two treatments, where the personal payment in both treatments is after a month, but in one treatment the charity payment is immediate whereas in the other it is after a month. If no difference in contributions is found, then that would support a warm glow explanation.

5 Behavioral Foundations

We examine here the behavioral foundations of our model of guilt. This allows us to confirm that the behaviors that characterize the model are peculiar to the existence of temptation and guilt, and this provides justification for interpreting our model as one of temptation and guilt. The purpose of this section is to provide a complete presentation of our model.\footnote{We claim no technical novelty here. The formal framework for studying temptation introduced by Gul and Pesendorfer [22] is now sufficiently well-developed in the decision theory literature to accommodate our theory of guilt (Noor [34], Kopylov [26], Kopylov and Noor [28]).}

As noted in Section 2, the agent’s choice of menus contains information about normative and temptation preferences over alternatives, which in turn contains information about anticipated choice from menus, and the anticipated costs of self-control and guilt. Therefore, assuming that the agent is sophisticated in the sense of being able to foresee his future choices and experiences, the identification of the normative and temptation preferences over
menus is sufficient to identify all the relevant components of our model. We exploit ideas developed in the axiomatic literature following GP [22]. As in Noor [34], we hypothesize that choices from a distance reflect the normative perspective. Instead of formulating an infinite horizon model as in Noor [34], we communicate the basic ideas in the simplest possible way here.\footnote{The ideas in [34] can be used to formulate an infinite horizon version of our model in which every period the agent picks current consumption and also a menu for the next period. The resulting model is such that each period integrates both the period 1 and period 2 choices. That is, the agent experiences guilt from immediate bad consumption, but is tempted to avoid future guilt and thus chooses menus accordingly.} Specifically, we simply assume that there is an ex ante stage – a “special period 0” – that precedes period 1, and that the agent’s choices in that period are fully normative. Thus guilt is absent in period 0, though anticipated guilt is not. While our terminology suggests that the ‘distance’ between periods 0 and 1 is temporal, other forms of distance may be invoked, such as if the agent was asked to dictate choices for others.

While the agent chooses alternatives in period 2 and menus in period 1, we take as a primitive his (normative) preferences over menus of menus in period 0. To visualize these objects, consider the following examples. If a restaurant is a menu $a$, then a neighborhood of restaurants is a menu of menus $A = \{a_1, a_2, \ldots\}$. Similarly if the allocation choices available in a dictator game is a menu $a$ and the exit option is a menu $b$ (yielding a single allocation), then a choice of whether to exit or play the game is a menu of menus $A = \{a, b\}$.

Our task is to (i) identify normative and temptation preferences over menus, and (ii) behaviorally characterize the functional forms (4)-(5). This is done through period 0 preference using ideas familiar from GP but extended to a setting where preferences are defined over menus of menus. So, if the agent prefers to commit to menu $a$ rather than $b$:

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

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. Moreover, if she exhibits a preference for commitment

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

then the preference for avoiding menu $b$ reveals that $b$ is tempting. These
connection permits us to take the next step and identify the behavioral implications of our model for \( \succeq \).

5.1 Preliminaries

For any compact metric space \( Z \), let \( \Delta(Z) \) denote the set of probability measures on the Borel \( \sigma \)-algebra of \( Z \), endowed with the weak convergence topology; \( \Delta(Z) \) is compact and metrizable [1], and we often write it simply as \( \Delta \) with generic elements \( x, y, z \), and we often refer to these as alternatives. Let \( \mathcal{M}_1 = \mathcal{K}(\Delta) \) denote the set of all nonempty compact subsets of \( \Delta \), 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. For \( \alpha \in [0, 1] \), \( \alpha x + (1 - \alpha)y \in \Delta \) is the \( \alpha \)-mixture that assigns \( \alpha x(S) + (1 - \alpha)y(S) \) to each \( S \) in the Borel \( \sigma \)-algebra of \( Z \). Similarly, \( \alpha a + (1 - \alpha)b \equiv \{\alpha x + (1 - \alpha)y : x \in a, y \in b\} \in \mathcal{M}_1 \) is an \( \alpha \)-mixture of menus \( a \) and \( b \). 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. Its generic elements are \( A, B, C \) and an \( \alpha \)-mixture of menus \( A \) and \( B \) is given by \( \alpha A + (1 - \alpha)B \equiv \{\alpha a + (1 - \alpha)b : a \in A, b \in B\} \in \mathcal{M}_0 \). Both \( \mathcal{M}_0 \) and \( \mathcal{M}_1 \) are compact (Aliprantis and Border [1, Theorem 3.71]) and the mixture operations in these spaces are continuous.

Our primitive is a preference relation \( \succsim \) on \( \mathcal{M}_0 \). 13 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. Choice in the ex ante stage is prior to the experience of temptation, and choice in the remaining stages is subject to temptation. In particular, 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 (behavioral assumptions) imposed on \( \succsim \) characterize the following basic representation theorem (see

---

13 This choice domain is used in Kopylov and Noor [28]. An important difference between this domain and the domain of multi-period menus used in GP [23] is the absence of lotteries over menus. This feature is obtained by exploiting Kopylov [26].
In the interest of brevity, we simply assert the existence of this representation as an axiom. Say that a function \( u: X \to \mathbb{R} \) is linear if for all \( \alpha \in [0, 1] \) and \( x, y \in X \),
\[
u(\alpha x + (1 - \alpha)y) = \alpha u(x) + (1 - \alpha)u(y).
\]
Let \( \mathcal{U} \) be the set of all continuous linear functions \( u: X \to \mathbb{R} \). Similarly, define linearity for functions on \( \mathcal{M}_1 \) and let \( \mathcal{U}_1 \) be the set of all continuous linear functions \( V: \mathcal{M}_1 \to \mathbb{R} \).

**Axiom 1 (Basic)** The preference \( \succsim \) has a utility representation \( U: \mathcal{M}_0 \to \mathbb{R} \) such that for all \( A \in \mathcal{M}_0 \) and \( a \in \mathcal{M}_1 \),
\[
U(A) = \max_{a \in A} \left( U(a) - \left[ \max_{b \in A} V(b) - V(a) \right] \right),
\]
\[
s.t. \ U(a) = \max_{x \in a} \left( u(x) - \left[ \max_{y \in a} v(y) - v(x) \right] \right),
\]
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 (6), which in turn identifies the agent’s normative and temptation preferences over menus, represented by \( U \) and \( V \) respectively. The desired functional form (7) for normative utility \( U \) is asserted. However, the desired functional form for \( V \), given by (5), has yet to be ascertained. We turn to this now.

### 5.2 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 menus $a, b \in \mathcal{M}_1$, 

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

Take any pair of singleton menus, $a = \{x\}$ and $b = \{y\}$. Temptation Consistency states that committed consumption $\{y\}$ is more tempting than $\{x\}$ in the interim stage if and only if $y$ is more tempting than $x$ ex post. 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.

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 correctly assessing ex post choice implies also a correct assessment of ex post guilt.

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

$$\{a \cup b\} \succ \{b\} \implies \{a\} \not\succ \{a, a \cup b\}.$$ 

The axiom states that if the agent anticipates that the ex post choice from $a \cup b$ is not an element of $b$, then in the interim stage the agent can never be tempted by the menu $a \cup b$, relative to the menu $a$. The bite comes in cases where $b$ contains something tempting. The axiom then says that the presence of that temptation does not attract the agent toward $a \cup b$ in the interim period, if it is the case that the temptation in $a \cup b$ is never chosen. The contrapositive of the axiom implies that if $a \cup b$ tempts $a$ in the interim stage, then it must be because $a \cup b$ contains overwhelming temptation.

The final axiom is the substantive one that relates temptation by menus to guilt. Define the relation $\succ_0$ over $\mathcal{M}_1$ by 

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

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

$$\{a \cup b\} \succ \{a \cup b, b\} \implies a \succ_0 b.$$
Suppose \( b \) contains greater ex post temptation than \( a \), in which case \( a \) and \( b \) offer the same ex post temptation. The axiom states that if \( b \) is more tempting than \( a \) in the interim stage, then it must be because consumption from \( b \) is subject to less guilt. This is reflected in the fact that \( b \) offers strictly less virtuous consumption than \( a \), that is, \( a \succ \succ b \).

Finally, say that \( \succ \) 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 \( \succ \) satisfies Basic, Temptation Sophistication, Temptation Consistency and Guilt-Averse Temptation if and only if \( \succ \) has a utility representation (6)–(7) such that for all \( a \in \mathcal{M}_1 \),

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

where \( \kappa > 0 \) and \( u, v \) are affinely independent.

Thus, under the three behavioral assumptions in addition to Basic, we obtain a characterization of our model in terms of ex ante preference. The proof is presented in the appendix. We close by confirming that our representation has the desired uniqueness properties.

**Theorem 2** If \( \succ \) is nondegenerate preference that admits a representation (6)–(8), 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 \( \beta_u, \beta_v \in \mathbb{R} \).

### 5.3 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, \( \succ \) and \( \succ^* \), 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 \( x, y \in \Delta(Z) \),

\[
\{\{x\}\} \succ \{\{y\}\} \iff \{\{x\}\} \succ^* \{\{y\}\}, \text{ and } \{\{x\}\} \succ \{\{x, y\}\} \iff \{\{x\}\} \succ^* \{\{x, y\}\}.
\]
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-ridden indulgence is experienced. To formalize this idea, say that

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

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

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 $a$. 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 $\succ$ despite this, then $a \cup b$ will also tempt the less guilt-prone agent $\succ^*$. The contrapositive states that if $a \cup b$ ‘normatively improves’ on $b$ to the extent that an agent $\succ^*$ ceases to be tempted by it, then the more guilt-prone agent $\succ$ will also cease to be tempted by it. The next theorem characterizes this definition.

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

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

(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

---

14Suppose 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.
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 – 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 $\succ^*$ is able to resist temptation, then so is $\succ$.

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

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

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

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

A curious feature of our model is its lack of parameters – there is only one, namely, $\kappa$. 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.

We conclude by noting that the parameter $\kappa$ can be interpreted in terms of how the agent discounts, in the interim period, 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 implies greater self-control in the interim stage.
6 Related Models

The GP [22] 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 \(\succ^1\) 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 [34]) 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-Sadowski [15] specialize to the dictator game setting, and study the shame associated with behaving unfairly 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 (that is, preferences when she is not observed by the recipient) and \(v\) reflects the perceived social norm. If the agent does not behave exactly as dictated by \(v\) – and in particular behaves ‘selfishly’ by following \(u\) – then she experiences shame. That is, the analog of ‘self-control costs’ here is ‘shame’. Unlike this paper, the authors’ motivation is mainly to model the exit decision by dictators, rather than other evidence from the literature. Moreover, the authors do not propose to explain social preferences in terms of temptation, but rather reinterpret a temptation model in a way that behaviorally reflects the social image theory of social preferences that has been studied in economics.

Kopylov [27] 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 extension of GP given by:

\[
W(x) = \max_{\mu \in x} \{u(\mu) - \left( \max_{\eta \in x} v(\eta) - v(\mu) \right) + k \left( \max_{\eta \in x} u(\eta) - u(\mu) \right) \}, \quad x \in Z.
\]

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.

Though not intended as such by the authors, the above models of painful emotions can be readily reinterpreted as models of guilt – in the first model
shame would be reinterpreted as guilt and in the second it would be the cost of deviating from perfectionist standards. 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. In particular, the reinterpreted models would be consistent with the example discussed in Section 2.2. That is, they would exhibit

$$\{B\} \succ^1 \{g, b\},$$

where $B$ is a normatively worse alternative than $b$. This would be the peculiar implication of guilt in these models: the agent would commit to something worse than $b$ in order to avoid the pain of guilt. While these reinterpreted models exhibit guilt aversion, our model provides also a theory of guilt aversion. Specifically, it hypothesizes that guilt-aversion is a temptation and that it is welfare-irrelevant.

For completeness, we observe that 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 differs from the two above. Most notably, while preferences over singleton menus in the above models is represented by $u$, in our model it is represented by $u + \kappa v$. The interpretation is that temptation acts on the agent even at the time of choice of menu, and in particular, even choice among singleton menus is a tug-of-war between normative and temptation motivations. As in the special case of Kopylov [27], there is a desire to avoid self-control costs and also guilt costs (these reduce utility). However our model makes explicit that the desire to avoid self-control costs is a feature of normative motivations, whereas the desire to avoid guilt is that of temptation. Formal justification for this requires richer data than a preference over menus – see Section 5.

7 Concluding Remarks

In this paper we observe that a study of guilt calls for a consideration of ‘guilt by guilt-aversion’. We build a model that treats guilt-aversion as a temptation, rather than as a welfare-relevant feature of normative preference. The model’s foundations give behavioral support to our interpretation of
the constructs used in the model. An analysis of guilt-proneness reveals highly intuitive properties of the model. This paper applies the model to the topic of social preferences and moral hypocrisy. The model is shown to accommodate a wide range of evidence that seems to suggest that people are moral hypocrites, and suggests different welfare implications than those implied by the view that people behave generously due to social pressure rather than a moral imperative. Our experiment suggests the existence of a moral imperative to be generous that struggles with a temptation to be selfish.

An interesting next step in the study of guilt is the study of self-deception. While this paper has shown that agents may take particular physical actions to alleviate guilt, psychologists have argued that the mind can take actions of its own to alleviate guilt. By selective memory and selective or biased interpretation, an agent may find ways to justify immoral behavior. Such justification helps them view the immoral behavior as moral, and thus to avoid guilt, at least temporarily. Guilt may still be borne ‘deep inside’ or at later points in time, but self-deception places some wedge between immoral action and the experience of guilt.\footnote{The reader is referred to Kopylov and Noor \cite{28} for an axiomatic model of self-deception (albeit without guilt).}

\section*{A Appendix: Proofs of Propositions}

\textbf{Proof of Prop 1.} Due to the hump-shape of normative utility \( u(M - s, s) \),

\[ \frac{du}{ds} = u_2 - u_1 \begin{cases} < 0 & \frac{M}{2} < s < 1 \\ > 0 & 0 < s < \frac{M}{2} \end{cases} \]

\[ \frac{d^2u}{ds^2} = u_{22} - 2u_{12} + u_{11} < 0 \quad s \in [0, M] \]

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

\[ \frac{dv}{ds} = v_2 - v_1 < 0 \]

\[ \frac{d^2v}{ds^2} = v_{22} - 2v_{12} + v_{11} < 0 \]
Dictators choose \( s \) 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 \( s \leq \frac{M}{2} \). 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 \( \lambda_0 \) be the threshold temptation intensity such that,

\[
\frac{du}{ds} + \lambda_0 \frac{dv}{ds} |_{s=0} = 0.
\] (9)

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

\[
\frac{du}{ds} + \lambda \frac{dv}{ds} |_{s=0} = 0 \\
\Rightarrow \frac{du}{ds} + \lambda \frac{dv}{ds} |_{s \in (0, \frac{M}{2})} < 0 \\
\Rightarrow s_\lambda = 0, \text{ that is, the agent continues not to share for any temptation intensity } \lambda \text{ higher than } \lambda_0. \text{ For } \lambda \text{ strictly lower than } \lambda_0, \text{ we have } s_\lambda \in (0, \frac{M}{2}), \text{ which is determined by the first order condition,}
\]

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

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

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

Therefore, we have that as \( \lambda \) increases from 0 to \( \lambda_0 \), \( s_\lambda \) decreases from \( \frac{M}{2} \) to 0. \( \blacksquare \)

**Proof of Prop 2.** Recall the threshold \( \lambda_0 \) defined in (9). In order for \( C(dg) = \{(M - s, s)\} \) for \( s > 0 \), it must be that \( \lambda < \lambda_0 \). Thus define \( \lambda^* = \lambda_0 \). For \( C_1(\{dg, e\}) = \{e\} \), we need to establish

\[
U(dg) + V(dg) < U(e) + V(e).
\]

Observe that

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

So the desired inequality requires

\[
2 [u(M - s\lambda, s\lambda) + \lambda v(M - s\lambda, s\lambda)] < u \left( \frac{M}{2}, \frac{M}{2} \right) + u(M, 0) + 2\lambda v(M, 0)
\]

Define the difference

\[
D(\lambda) := U(dg) + V(dg) - [U(e) + V(e)]
\]

\[
= 2u(M - s\lambda, s\lambda) - u \left( \frac{M}{2}, \frac{M}{2} \right) - u(M, 0) + 2\lambda [v(M - s\lambda, s\lambda) - v(M, 0)].
\]

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 \( \lambda \), and observe

\[
\frac{dD(\lambda)}{d\lambda} = \frac{2du}{ds} \frac{ds\lambda}{d\lambda} + \frac{2\lambda dv}{ds} \frac{ds\lambda}{d\lambda} + 2 \frac{v(M - s\lambda, s\lambda) - v(M, 0)}{ds} = 0
\]

\[
= 2 [v(M - s\lambda, s\lambda) - v(M, 0)] < 0
\]

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

**Proof of Prop 3.** Begin by noting that \( U(a) + V(a) = U(\{M - c, 0\}) + V(\{M - c, 0\}) \) implies

\[
2u(M - s\lambda, s\lambda) - u \left( \frac{M}{2}, \frac{M}{2} \right) - u(M - c, 0) + \lambda [2v(M - s\lambda, s\lambda) - v(M, 0) - v(M - c, 0)] = 0
\]

\[
\Rightarrow \ u_1(M - c, 0) dc + \lambda v_1(M - c, 0) dc + [2v(M - s\lambda, s\lambda) - v(M, 0) - v(M - c, 0)] d\lambda = 0
\]

\[
\Rightarrow \ dc = \frac{v(M, 0) + v(M - c, 0) - 2v(M - s\lambda, s\lambda)}{u_1(M - c, 0) + \lambda v_1(M - c, 0)}.
\]

Recall the quantities \( \lambda_0 \) and \( D(\lambda) \) defined in the proofs of earlier propositions. Then \( s\lambda = 0 \) whenever \( \lambda \geq \lambda_0 \), and moreover,
\[ D(\lambda) > 0 \]
\[ \implies U(dg) + V(dg) < U(e) + V(e) \]
\[ \implies c > 0. \]

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 \( \frac{dc}{d\lambda} \bigg|_{\lambda = \lambda_0} < 0 \). By continuity, there exists \( \lambda^{**} < \lambda_0 \) such that \( \frac{dc}{d\lambda} < 0 \) for all \( \lambda \geq \lambda^{**} \).

**Proof of Prop 4.** Given our assumptions on \( u \) and \( v \), we have \( u(5, 5) > u(6, 1) \) and \( v(6, 1) > v(5, 5) \). Note that

\[
\begin{align*}
U(a_1) + V(a_1) &= u(6, 5) + \lambda v(6, 5) \\
U(a_2) + V(a_2) &= \max\{2[u(6, 1) + \lambda v(6, 1)], 2[u(5, 5) + \lambda v(5, 5)]\} - [u(5, 5) + \lambda v(6, 1)] \\
U(a) + V(a) &= u(6, l) + \lambda v(6, l) = \frac{1}{2}[u(6, 5) + u(6, 1)] + \frac{1}{2}\lambda[v(6, 5) + v(6, 1)]
\end{align*}
\]

Let \( \lambda^{***} \) 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, \( \lambda^{***} \) 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,

\[
\begin{align*}
U(a) + V(a) &= U(\frac{1}{2}, a_1; \frac{1}{2}, a_2) - V(\frac{1}{2}, a_1; \frac{1}{2}, a_2) \\
&= \frac{1}{2} \left\{ u(6, 5) + u(6, 1) + \lambda v(6, 5) + \lambda v(6, 1) + u(5, 5) \right\} \\
&= \frac{1}{2} \left\{ u(5, 5) - \lambda v(6, 5) - 2u(6, 1) - \lambda v(6, 1) \right\} \\
&> 0
\end{align*}
\]

Thus \( C_1(\{a, (\frac{1}{2}, a_1; \frac{1}{2}, a_2)\}) = \{a\} \). ■

**B Appendix: Experiment Instructions**

**B.1 Treatment A**

Please read the instructions carefully. Failure to follow them will invalidate your responses, and no payments will be made.

**Your Task.** There are $10 available to you at the end of this class, and you will divide this between yourself and the charity *American Red Cross*. You can keep it all or keep nothing, or keep part of it. The only restriction is that the smallest denomination should be 25 cents.
You will receive your payment *in cash* at the end of *this* class, and the charity will be sent its payment electronically over the internet within the hour following the class.

**How.** (a) You have up to 10 minutes to decide how to divide the $10 between yourself and the charity. Write your share (in dollars) on the answer sheet and inside the flap of the envelope.\(^\text{16}\)

(b) Write your *study code* on the answer sheet and on the envelope. Your study code is the first two letters of mother’s first name, the first two letters of father’s first name, the first two letters of month of birth and the last 2 digits of your social security number.\(^\text{17}\)

(c) Fold the answer sheet and put it in the envelope. Do *not* seal the envelope. Hand in the envelope.

**Protocol for your Payment.**
At the end of class, all subjects will be asked to approach the desk one by one. Tell the cashier your study code, and you will receive your envelope, sealed with your payment inside.

**Confidentiality.** It is crucial for this experiment that your individual allocation decision be confidential. Do *not* write your name or student ID anywhere on the answer sheet. Keep your answer sheet hidden from *everyone* around you. The only person who will ever see your allocation decision will be the cashier, *but s/he will not know your identity*! The Principal Investigators will only see anonymous responses.

### B.2 Treatment B

Please read the instructions carefully. Failure to follow them will invalidate your responses, and no payments will be made.

**Your Task.** There are $10 available to you after a month, and today you will divide this future amount between yourself and the charity *American Red Cross*. You can keep it all or keep nothing or keep part of it.

\(^{16}\)If you write different amounts you will not receive any payment.

\(^{17}\)E.g. if your mother’s and father’s first names are Jane and John resp., your birthday is in Oct and the last 2 digits are 22, then your study code is JAJOOC22.
You will receive your payment *in cash* at the end of class on 2nd Aug 2010, and the charity will be sent its payment electronically over the internet within the hour following that class.

**How.** (a) You have up to 10 minutes to decide how to divide the $10 between yourself and the charity. Write *your* share (in dollars, smallest denomination 25 cents) on the answer sheet.

(b) To preserve complete confidentiality, you need a *code*: it is constructed by taking the first two letters of mother’s first name, the first two letters of father’s first name, the first two letters of month of birth and the last 2 digits of your social security number.\(^{18}\) Write your code on the answer sheet and on the envelope.

(c) Fold the answer sheet and put it in the envelope and hand it in. Do *not* seal the envelope.

**Protocol for your Payment.**

At the end of class on the 2nd of Aug 2010, you will tell the cashier your code and subsequently receive your envelope, sealed with your payment inside.\(^{19}\) The cashier will not have any way of knowing the amount of payment you are collecting.

**Confidentiality.** It is crucial for this experiment that your individual allocation decision be confidential. Do *not* write your name or student ID anywhere on the answer sheet. Keep your answer sheet hidden from *everyone* around you. Your identity will not be known by the cashier when your payment is being placed in the envelope, and the amount of payment will not be known by the cashier when you collect the sealed envelope. The Principal Investigators will only see anonymous responses.

### C Appendix: Proof of Theorem 1

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

\[
V(a) = \kappa \max_a (u + v) + k_2 \max_a v + k_3 \max_a u
\]  

\(^{18}\)E.g. if your mother’s and father’s first names are Jane and John resp., your birthday is in Oct and the last 2 digits are 22, then your study code is JAJOOC22.

\(^{19}\)If you do not collect your envelope that day we will help you get your payment as soon as possible. The charity will be paid as scheduled.
Proof. We first show that
\[
\max_a u = \max_b u \text{ and } \max_a (u + v) = \max_b (u + v) \text{ and } \max_a v = \max_b v \implies V(a) = V(b).
\]
(11)
This is proved as in Kopylov and Noor [28]. 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 \(\succ\) is nondegenerate, there exist \(x^*, y^*\) such that \(U(\{x^*\}) > U(\{x^*, y^*\}) > U(\{y^*\})\). Let \(a^* = \varepsilon \{x^*\} + (1 - \varepsilon) a\) and \(b^* = \varepsilon \{y^*\} + (1 - \varepsilon) b\) and take \(\varepsilon > 0\) s.t.
\[
(U + V)(a^*) < (U + V)(b^*).
\]
(12)
Such \(\varepsilon\) exists by continuity of \(U, V\). Note also that by linearity of \(U, V\) and \(u\),
\[
U(a^*) > U(a^* \cup b^*) > U(b^*)
\]
(13)
and
\[
\max_{a^*} u < \max_{b^*} u.
\]
(14)
Given (12), there are two possibilities.

- \((U + V)(a^*) < (U + V)(a^* \cup b^*)\). Combined with (13), \(V(a^*) < V(a^* \cup b^*)\). Therefore, \(\{a^*\} \succ \{a^*, a^* \cup b^*\}\), which contradicts Temptation Sophistication.

- \((U + V)(b^*) > (U + V)(a^* \cup b^*)\). Combined with (13), \(V(b^*) > V(a^* \cup b^*)\). Therefore, \(\{a^* \cup b^*\} \succ \{a^* \cup b^*, b^*\}\). However, given (14), this contradicts Guilt-Averse Temptation.

This establishes (11). 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 (11) yields the desired form for \(V\) on \(\mathcal{M}_1^c\) by an application of Harsanyi’s aggregation theorem (see Border [8]).

20 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.
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:

$$
\begin{align*}
u(x') &= u(y'), v(x') < v(y') \\
v(x'') &= v(y''), u(x'') > u(y'')
\end{align*}
$$

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

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

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

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

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

implies $(\kappa + k_3)u(x'x'') + (\kappa + k_2)v(x'x'') > (\kappa + k_3)u(y'y'') + (\kappa + k_2)v(y'y'')$, then we find $\theta$ such that $f(\theta) < \frac{\kappa + k_2}{\kappa + k_3}$ and use an analogous argument to show that $V(\{x'x''\}) > V(\{y'y''\})$.

Next, observe that the linearity of $u$ and $v$ implies

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

and similarly, $v(x'x'') < v(y'y'')$. Letting $x := x'x''$ and $y := y'y''$, 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 $\{x\} \sim \{x\}$, 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 $V$,
\[
V(\{x\}) = (\kappa + k_3) u(x) + (\kappa + k_2) v(x) < (\kappa + k_3) u(y) + (\kappa + k_2) v(y) = V(\{y\}).
\]
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, $\{\{x^*\}\} \succ \{\{x^*, y^*\}\} \text{ and } \{\{x^*\}\} \sim \{\{x^*, y^*\}\}$, contradicting Temptation Consistency. Thus, $\kappa + k_2 > 0$. This completes the proof. \[ \blacksquare \]

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

Lemma 4 The representation $(u, v, \kappa)$ is unique in the sense that $(u', v', \kappa')$ 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 requirement is 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. \[ \blacksquare \]
D  Appendix: Proof of Theorems 3 and 4

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

Lemma 5 $\succsim$ 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,

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

for some $w, z$. Observe that by the representation, for all $\alpha$,

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

However, there exists $\alpha$ such that

$$\{\alpha w + (1 - \alpha) x\} \nless \{\alpha w + (1 - \alpha) x, \alpha z + (1 - \alpha) y\},$$

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 $\lambda$ if necessary. ■

Lemma 6 $\succsim$ 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:

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

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

Thus, $u = \lambda u^*$ for some $\lambda > 0$. Hence, $\lambda < 1$. Since $\succsim$ and $\succsim^*$ are nondegenerate, we can assume $u(x') < u(y')$ and $v(x') = v(y')$ wlog. Then, $\{x\} \succ \{y\}$ and $\{x\} \nless \{y\}$, a contradiction. Therefore, $\lambda \geq 1$.

$\Longleftarrow$: Suppose $\lambda \geq 1$. Then, $u(x) = \lambda u(y)$, $v(x) = \lambda v(y)$ for some $x, y$. By nondegeneracy and ex post similarity, $u(x') = \lambda u(y')$ and $v(x') = \lambda v(y')$. Thus, $\{x\} \nless \{y\}$, a contradiction. Therefore, $\lambda < 1$. Since $\succsim$ and $\succsim^*$ are nondegenerate, we can assume $u(x') < u(y')$ and $v(x') = v(y')$ wlog. Then, $\{x\} \succ \{y\}$ and $\{x\} \nless \{y\}$, a contradiction. Therefore, $\lambda \geq 1$. ■
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 $\theta$ such that
\[
\lambda < f(\theta) < 1.
\]
Let $x = x'\theta x''$, $y = y'\theta y''$ and $z = z'\theta z''$. Then,
\[
\begin{align*}
f(\theta) < 1 & \implies \frac{u(x) - u(y)}{v(y) - v(x)} < 1 \implies u(y) + v(y) > u(x) + v(x), \quad (15) \\
f(\theta) > \lambda & \implies \frac{u(x) - u(y)}{v(y) - v(x)} > \lambda \implies \frac{u(y)}{\lambda} + v(y) < \frac{u(x)}{\lambda} + v(x), \quad (16) \\
u(x) > u(y), v(x) < v(y) & \quad (17)
\end{align*}
\]
Let $a = \{x\}$, $b = \{y\}$. Then $a \cup b = \{x, y\}$. By (17),
\[
\{a\} \succ \{a \cup b\}, \{a\} \succ^* \{a \cup b\} \quad (18)
\]
By (15),
\[
\begin{align*}
V(a \cup b) &= \kappa [u(y) + v(y) - u(x)] \succ \lambda v(x) \quad (19) \\
V(a) &= \kappa [u(x) + v(x) - u(x)] = \lambda v(x) \quad (20)
\end{align*}
\]
This implies, together with (18), $\{a\} \succ \{a, a \cup b\}$.

On the other hand, by (16),
\[
\begin{align*}
V^*(a \cup b) &= \kappa^* \left[ \frac{u(x)}{\lambda} + v(x) - \frac{u(x)}{\lambda} \right] = \kappa^* v(x) \\
V^*(a) &= \kappa^* \left[ \frac{u(x)}{\lambda} + v(x) - \frac{u(x)}{\lambda} \right] = \kappa^* v(x)
\end{align*}
\]
\[
\implies V^*(a \cup b) = V^*(a) \quad (21)
\]
Together with equation (18), $\{a\} \sim^* \{a, a \cup b\}$, which contradicts with that $\succ^*$ is more guilt-prone then $\sim^*$.

\[\iff: \text{If } \lambda = 1 \text{ then the proof is trivial. So suppose } \lambda > 1.\text{The preference } \{a\} \succ \{a, a \cup b\} \text{ implies,}
\]
\[
\begin{align*}
U(a) &> U(a \cup b) \implies u(x) > \max_{a \cup b} \left[ u + \left( v - \max_{a \cup b} v \right) \right] \\
V(a) &< V(a \cup b) \implies v(x) < \max_{a \cup b} \left[ v + \left( u - \max_{a \cup b} u \right) \right]
\end{align*}
\]

39
By these inequalities and by definition of $V^*$,

$$V^* (a \cup b) = \max_{a \cup b} \left[ v + \frac{1}{\lambda} \left( u - \max_{a \cup b} uv \right) \right]$$

$$\geq \max_{a \cup b} \left[ v + \left( u - \max_{a \cup b} u \right) \right]$$

$$\geq \max_{a \cup b} v (x) = V^* (a)$$

On the other hand, by the definition of $U^*$,

$$U^* (a \cup b) = \max_{a \cup b} \left[ \frac{1}{\lambda} \left( u + \lambda \left( v - \max_{a \cup b} v \right) \right) \right]$$

$$\leq \max_{a \cup b} \left[ u + \left( v - \max_{a \cup b} v \right) \right]$$

$$< \frac{1}{\lambda} u (x) = U^* (a)$$

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

**Lemma 7** $\lambda \geq 1$ if and only if $\succsim$ has more ex post self-control than $\succsim^*$.

**Proof.** $\implies$: 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, $\{\{x\}\} \succ \{\{x, y\}\} \sim \{\{y\}\}$.

$\impliedby$: 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 $f (\theta) \in (1, \frac{1}{\lambda})$). That is,

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

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

a contradiction. ■
References


