Conscience Accounting: Emotional Dynamics and Social Behavior

Uri Gneezy[*],

Alex Imas[†]

Kristóf Madarász[‡]

February 25, 2012

Abstract

We develop a dynamic model where people decide in the presence of moral constraints and test the predictions of the model through two experiments. Norm violations induce a temporal feeling of guilt that depreciates with time. Due to such fluctuations of guilt, people exhibit an endogenous temporal inconsistency in social preferences—a behavior we term conscience accounting. In our experiments people first have to make an ethical decision, and subsequently decide whether to donate to charity. We find that those who chose unethically were more likely to donate than those who did not. As predicted, donation rates were higher when the opportunity to donate came sooner after the unethical choice than later. Combined, our theoretical and empirical findings suggest a mechanism by which prosocial behavior is likely to occur within temporal brackets following an unethical choice.

JEL classification: D03, D06, D09

Keywords: Emotions, Temporal Brackets, Deception, Prosocial Behavior

[*]Rady School of Management, University of California, San Diego, La Jolla, CA 92093
[†]Dept. of Economics, University of California, San Diego, La Jolla, CA 92093
[‡]London School of Economics, London WC2A 2AE, United Kingdom; k.p.madarasz@lse.ac.uk

1 Introduction

> Terrible is the Temptation to do Good – Bertolt Brecht, *The Caucasian Chalk Circle* (1944)

In this paper, we report the results of two experiments in which people who first made an unethical choice were then more likely to donate to charity than those who did not. We interpret these results in the context of our model which explores how retrospective emotions impact social behavior. We focus on the emotion of guilt and demonstrate how dynamic fluctuations of guilt induced by past unethical behavior act as a motivator for prosocial behavior, as well as a potential deterrent from norm violations. We term this effect conscience accounting.

Our contribution is twofold. First, we offer a simple model where emotional fluctuations triggered by past decisions induce a temporal shift in preferences, and individuals take these effects into account ex ante. Second, we present a novel experimental paradigm in which we can directly test and identify the effects of emotions, specifically guilt, on choice behavior.

Throughout history, institutions have been built to take advantage of the effects of guilt on charitable behavior and to enable individuals to account for their conscience. The medieval Catholic Church's practice of granting "indulgences" absolved an individual of sins through a system of "tariff penances," whereby a particular amount of money transferred to the Church would pardon an individual from certain sins. Sins were priced on a sliding scale according to which serious sins were more expensive to pardon than smaller ones, and the Church used groups of professional "pardoners," or quaestores, to collect money from willing individuals. Today, Mass in the Catholic Church typically involves congregants reciting a prayer called the Confiteor in which they confess, and are in turn reminded of, their sins. A collection plate is passed around afterwards to solicit alms. The Catholic Church was not alone in the institutionalization of conscience accounting. Around the time of the Second Temple—500 B.C. to 70 A.D.—Jewish leaders formalized the use of chatot (sin) and ashamot (guilt) offerings as atonement for transgressions. Individuals made these offerings through the purchase of korban—an animal

sacrifice only the Temple priests could perform. Different sins required different levels of sacrifice. Vendors around the Temple sold doves for trivial sins and lambs for those considered more damning.

These kinds of institutions imply some form of moral constraints that people impose on themselves. Recent experimental findings support this observation in giving environments (e.g., Dana, Weber, and Kuang 2007; Lazear, Malmendier, and Weber 2011). In other moral choices, such as the decision to deceive, research has shown that people have an associated internal moral cost that manifests itself as a conditional aversion to lying (e.g., Gneezy, 2005; Sutter, 2009; Dreber and Johannesson, 2008).

We begin with the observation that violations of moral constraints are costly in terms of the guilt they induce, and develop a model of dynamic emotional decision making. The decision maker is subject to emotional fluctuations: certain actions quickly change her emotional state, which—absent further stimuli—gradually reverts back to normal. Specifically, our decision maker experiences the emotion of guilt that arises after she violates an internalized norm or acts in a way that she views as unethical.[1] Key to our model is that guilt is not just an aversive feeling that decreases the decision maker's utility, but one which decreases the extent to which she cares about improving her own consumption relative to the consumption of others.

Consistent with the dynamic nature of emotions described by Elster (1998), we allow for moral debt to depreciate over time, such that after the initial endogenous increase in guilt, the decision maker's emotions revert back to their initial "cold" state. Thus our model examines the evolution of guilt in a dynamic framework that permits us to derive novel results on the intertemporal aspects of prosocial decision making.

The identification of our model relies on the fact that emotional fluctuations cause time inconsistency in the decision maker's behavior. After violating an internalized moral constraint, the individual experiences feelings of guilt that create an emotional bracket whereby the prosocial reversal in behavior is largest right after the norm violation and diminishes over time. We term this emotional response conscience accounting. Being at least partially aware of such time inconsistency, individuals may value commitment that would "tie their hands" against being

[1]Guilt is associated with moral transgressions (e.g., Baumeister, Stillwel and Heatherton, 1994), and the desire to avoid guilt has been established as equilibrium behavior within a game theoretic framework (Battigalli and Dufwenberg, 2007, 2009; see Dufwenberg and Gneezy, 2000, and Charness and Dufwenberg, 2006, for experimental evidence of guilt aversion). Guilt is also considered an aversive feeling that discourages norm violations (e.g., Akerlof and Kranton, 2000).

too generous after violating a norm, have a distinct preference for a *delayed choice* until guilt subsides, or absent such commitment, may avoid violating a norm altogether.

We test the main predictions of the model using two experimental paradigms. We find support for the prediction of conscience accounting: individuals who achieved a given payoff by deception or stealing were more likely to donate to charity than those who achieved the same payoffs in a more ethical manner. In addition, we find that this effect occurs within a temporal bracket where the increase in prosocial behavior is greatest directly after the unethical act and decreases with the passage of time.

Our results have a direct application to charitable contributions and volunteering behavior, suggesting an additional explanation for why people donate their money and time. Charity as both a virtue and an institution has been a prominent facet of civilization as far back as the public dispensaries of ancient Greece and the Charity temple on Rome's Capitoline Hill. Today, more than two-thirds of Americans make annual donations to charity and many engage in volunteer work. The willingness to give has puzzled economists for decades; not only because it contradicts the assumption that people are fueled solely by self-interest, but because it does not seem to be driven by one simple alternative (Becker, 1976; Vesterlund, 2003; Andreoni, 1990, 1995; Meier, 2007; DellaVigna, List, and Malmendier, 2011). Our model of emotional dynamics could help in explaining this phenomenon and provide a mechanism through which firms and organizations wishing to maximize contributions – such as airlines collecting money for carbon offsets – can use guilt efficiently as a motivator.

Our approach is linked to the economic literature of incorporating procedural norms into economic behavior, for example, Kahneman, Knetsch and Thaler (1986) and Akerlof and Kranton (2000, 2005). In our model, individuals would prefer to adhere to procedural norms when attaining a given consumption vector. Upon violating a norm, however, they exhibit a temporal altruistic preference-reversal toward others. In this manner our theory helps identify norm violations in observable behavior. Furthermore, this mechanism offers predictions on how a person's ability to compensate for norm violations ex-post changes her propensity to violate a norm in the first place.

In addition, the insights from our theory can be generalized to other emotions. For example, angering situations can be seen to cause a similar time inconsistency in behavior, where individuals are more likely to hurt and lash out at others within a temporal bracket directly after

being treated unfairly, and less likely to do so after having some time to "cool off." Being aware of this time inconsistency, individuals may value commitment that would allow them to delay their future responses and increasing their willingness to enter otherwise advantageous angering situations that constrict their ability to retaliate until they cool off.

Our theory contributes to the small body of work in economics that considers the role of emotions in behavior. For example, Loewenstein (1987) studies the role of anticipation on time preferences and Kőszegi and Rabin (2007) study the impact of prospective gain-loss utility relative to endogenous expectations on risk attitudes. Our approach differs from these models in that they focus on the impact of emotions on behavior before the resolution of some event, while we study *retrospective* emotions where the direct effects of emotions on behavior after the resolution of an event and also because we focus on social preferences. In such a domain, Card and Dahl (2011) provide evidence that the realization of unexpected losses in football matches provoke a quick increase in family violence around the end of the game—an effect which disappears soon thereafter.

The rest of the paper is organized as follows. In Section 2, we introduce the model, outlining the dynamics of emotion and their effect on preferences. In Section 3, we present evidence from a deception game experiment in which we test several of the main propositions. Section 4 lays out the results of an "over-paying" experiment that provides further support for the theory. In Section 5, we discuss several examples of how conscience accounting can be utilized by firms to maximize revenue, and posit how our theory can be generalized to other emotions.

2 Model

Emma faces a temporal sequence of allocation decisions (dictator games) before a final period of consumption. Examples of such decisions abound: sharing profits with a business partner, contributing to a social cause, taking on household duties. In each decision round t, Emma chooses a payoff (consumption) vector $\pi_t = (\pi_t(a), \pi_t(b))$, where the first component refers to her own payoff and the second to the payoff of the person with whom she interacts with, from a compact set of feasible payoff vectors $\Pi_t \subset \mathbb{R}^2$.[2] The final allocation is the sum of all chosen

[2]Although Emma may interact and care about numerous others, for simplicity we consider two-dimensional allocation spaces, where preferences can be expressed as a function of some aggregated payoff received by others – such as another person, members of a particular community, beneficiaries of the church etc. – which is monotone

allocations over T decision rounds, $\pi = \sum_{t=1}^{T} \pi_t \in \mathbb{R}^2$, and is consumed in the final round T. The decision environment can thus be summarized by $\Gamma = \{\Pi_t\}_{t=1}^{T}$.

2.1 Preferences

In specifying Emma's preferences, we extend the standard model of altruism in two ways. First, her utility from a final allocation depends on her emotional state, $d \in \mathbb{R}^+$, which we call *moral debt* and interpret as the intensity of her guilt. Second, she derives utility not only from the consumption of the final allocation, but also from the anticipation of this consumption event. In each decision round t, she derives anticipatory utility from her expectation of the final consumption vector as a function of her emotional state in that period d_t.

Emma's guilt is determined by whether she acts in accordance with her internalized moral constraints. As in the literature discussed before, we interpret moral constraints (norms) as internalized prescriptions against particular behavior. Moral constraints describe what Emma *should* not do, and hence these need not prohibit specific payoff allocations, but rather ways in which these allocations are attained. Examples of such procedural fairness include attaining the same payoff allocation by either lying or telling the truth, by stealing from business partners or receiving a gift, and having a clear preference one way or another.

Importantly, for the purposes of the model we do not need to specify the content of Emma's moral constraints. It suffices to partition the choice set Π_t into two subsets by letting $N_t \subset \Pi_t$ be the set of allocations that can only be attained by violating a moral constraint. We assume $\Pi_t \backslash N_t$ to be non-empty and to contain $(0, 0)$ whenever it is in Π. Although we take the set of moral constraints to be exogenous, our model will provide a mechanism that can help identify norm violations in dynamic choice situations. We discuss this in more detail at the end of this Section.[3]

As is typical of many emotions, a class of events triggers a rapid change in an individual's emotional state that is often increasing in the size of the stimulus. With time the emotional state reverts back to its unaroused state. As such, we make two general assumptions about the dynamics of Emma's moral debt d_t: **(i)** a norm violation committed in round t leads to an

in each underlying payoff component.

[3]Here we consider problems with perfect information. One can extend the framework to the case with uncertainty about the moral character of actions. If guilt is increasing in Emma's certainty that she violated a norm, our model will imply a similar information aversion as proposed by Rabin (1995) to identify norms.

increase in moral debt by round $t+1$, and **(ii)** existing moral debt gradually depreciates with time.

The following example describes the evolution of moral debt given a choice of π_t at time t:

$$d_{t+1} = \gamma d_t + \max\{\pi_t^m(b) - \pi_t(b), 0\}, \tag{1}$$

where $\gamma \in (0,1)$ and $\pi_t^m(b) = \sup \pi_t'(b)$ where supremum is taken over the set of payoff efficient allocation that belong to $\Pi_t \backslash N_t$. In words, after a norm violation, guilt increases in proportion to how much harm an unethical action causes others relative to Emma's most selfishly efficient, but still ethical, allocation choice.[4] We emphasize that the predictions derived in this paper do not depend on the details of the above specification. Along with assumptions **(i)** and **(ii)** above, it suffices to assume that the jump in moral debt after a norm violation is increasing in the payoff difference between an efficient reference payoff and what the other party receives. Hence the functional form assumptions above play no role in the analysis.

The shape of Emma's anticipatory utility is identical to her consumption utility. Formally, at any round t, Emma experiences instantaneous utility based on her expectation of the final allocation and the intensity of her guilt d_t. Thus, in round t, she experiences utility in the following form:

$$u_t = E_t u(\pi, d_t), \tag{2}$$

where $u : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$, $u \in C^2$ and u is strictly increasing and strictly concave in each payoff argument. In addition, $u_{\pi_a,\pi_b} > 0$, and thus preferences are convex conditional on the state.[5] Finally, we assume that the relevant boundary condition holds such that Emma never wants to end up with a non-positive own consumption.

The predictive capacity of the model stems from two assumptions on the impact of moral debt. First, guilt is an aversive emotion and therefore moral debt d_t is an economic bad. Second, the guilt is a substitute of own consumption and a weak complement for the consumption of others. Guilt is thus not simply a negative emotion, but one that decreases "self-love" relative to the love of others. Formally,

[4] In a game theoretical model of prospective guilt aversion, Battigalli and Dufwenberg (2007) make a similar assumption.

[5] The assumption that monetary payoffs are economic goods is consistent with the findings of Charness and Rabin (2002).

Condition 1 *For all* (π, d), $u_d < 0$, $u_{d,\pi(a)} < 0$ *and* $u_{d,\pi(b)} \geq 0$. *Furthermore,* $\lim_{\pi(a)\to 0} u_{\pi(a)} = \infty$.

To complete the description of Emma's preferences, as is standard, we posit that at any period t, she maximizes the sum of her anticipatory and consumption utilities:

$$U_t = E_t \sum_{s=t}^{T} u(\pi, d_s) \tag{3}$$

For simplicity, we set the usual discount factor to unity, but a lower discount factor or assigning different constant positive weights for different rounds would not change the model's qualitative results.

2.2 Dynamic Plans

Emotional fluctuations in our model imply that Emma's preferences over the set of final allocations will change over time. For example, Emma's preferences before acting immorally are less sensitive to guilt than in the period after the norm violation. Given such *time-inconsistency*, to solve the model we need to specify Emma's strategy at each round t. Since the problem is separable in time, let $s_t : H_{t-1} \to \Pi_t$ be Emma's strategy in round t, where H_{t-1} is the set of histories leading up to round t. Let her plan be a collection of such functions: $s = \{s_t\}_{t=1}^{T}$. Let $\overline{s}_{t+1}(h_t)$ denote her complete continuation strategy in s following history $h_t \in H_t$.[6]

Definition 1 *A plan* s^* *is optimal if it is feasible in* Γ, *and for any* t *and* $h_{t-1} \in H_{t-1}$, *it is true that for all* $\pi'_t \in \Pi_t$,

$$U_t(h_{t-1}, s_t^*(h_{t-1}), \overline{s}_{t+1}^*(h_{t-1}, s_t^*(h_{t-1}))) \;\geq\; U_t(h_{t-1}, \pi'_t, \overline{s}_{t+1}^*(h_{t-1}, \pi'_t)) \;,$$

In words, in an optimal plan Emma maximizes her preferences at each round given her past behavior and her rational expectations about her continuation strategy. Note that the since the space of histories is compact and utilities are continuous, it follows from Harris (1985) that an optimal solution exists. Since we assume sequential moves and perfect information, we consider only pure strategies.

[6]Note that while a relevant history at time t is given by $h_{t-1} = \{\pi_s, d_t\}_{s=1}^{t-1}$ in effect we can simplify this since once d_t is given only the sum $\sum_{s=1}^{t-1} \pi_s$ which matters.

When deriving the implications of the model below, we will impose the following monotonicity assumption on norm violations: if achieving an allocation from a set requires a norm violation, then more selfish allocations can only be attained through a norm violation. For example, if Emma needed to steal to earn \$100 and allocate \$30 to others, she cannot achieve an allocation of \$120 for herself and \$20 to others without stealing.

Condition 2 *Suppose $\pi_t \in N_t$. If π_t' is such that $\pi_t'(a) \geq \pi_t(a)$ and $\pi_t'(b) \leq \pi_t(b)$, then $\pi_t' \in N_t$.*

2.3 Predictions

To derive the implications of the model, it suffices to consider problems with two general compact linear budget sets containing the origin and three periods. With a slight abuse of notation let these two sets be $\Pi \subset \mathbb{R}^+ \times \mathbb{R}^+$ and $G \subset \mathbb{R}^- \times \mathbb{R}^+$. Since Emma can never obtain a strictly positive payoff from G we interpret this set as a pure donation set. We normalize the slope of Π to be 1 and denote the slope of G by $p_G \in (0, \infty)$.

Our first result identifies a weak form of conscience accounting in observable behavior. A potential norm violation in Π_1 is followed by a surprise option to donate from G. Specifically, when choosing from Π, Emma is unaware that she will be presented with the choice set G. We compare behavior across two scenarios: the donation set G follows the initial, potential norm violation either sooner or later.

Proposition 1 *Consider $\Gamma_{\widetilde{h}} = \{\Pi, G, \emptyset\}$ and $\Gamma_{\widetilde{c}} = \{\Pi, \emptyset, G\}$. It follows that $\pi^*_{\widetilde{c}}(a) \geq \pi^*_{\widetilde{h}}(a)$ and $\pi^*_{\widetilde{c}}(b) \leq \pi^*_{\widetilde{h}}(b)$.*

If Emma refrains from violating a norm initially, she experiences no increase in her moral debt. Hence her preferences over the final allocation in rounds 2 and 3 should be identical. In contrast, if she violates a norm in round 1, her moral debt rises as a result, and by round 2 she experiences guilt. Given how debt effects marginal utilities, the optimality of the round 1 choice implies that in round 2, Emma would like to re-allocate payoffs from herself to others. Emma's round 3 preferences only take into account her feelings in round 3, when her guilt is always lower than in round 2. Hence, she donates a greater fraction of her wealth in a "hot" state closer to the norm violation than in a "cold" state further away.[7]

[7]The above result demonstrates that fluctuations of guilt can cause Emma to choose dominated payoff allocations whenever $p_D > 1$. Importantly, the above result extends to the case where Π and D are discrete.

We now turn to a strong form of conscience accounting. Consider the same setup as before, but now assume Emma knows in advance that a donation option will be presented in the future, as well as when this option will be available. A sufficient condition for our next result is that transferring payoffs in Π is weakly more efficient than doing so in G, and hence donations per se do not involve efficiency gains.

Proposition 2 *Consider* $\Gamma_h = \{\Pi, G, \emptyset\}$ *and* $\Gamma_c = \{\Pi, \emptyset, G\}$. *If* $p_G \geq 1$, *it follows that* $\pi_c^*(a) \geq \pi_h^*(a)$ *and* $\pi_c^*(b) \leq \pi_h^*(b)$.

By virtue of rational expectations, Emma understands that her round 2 preferences are more affected by guilt than her round 3 preferences. Because transferring payoffs is more efficient in Π than in G, positive donations arise only as a mechanism of costly conscience accounting —brought about by temporal preference reversals due to the fluctuations of guilt. Hence, in an optimal plan, Emma internalizes the extent to which she will be too "tempted" to subsequently donate relative to her round 1 preferences. Since this temptation is weakly greater in Γ_h than in Γ_c, she is overall more altruistic in the former than in the latter.

A sufficient condition for Proposition 2 was that donations did not represent efficiency gains per se. If $p_G < 1$, this result need not hold. To see the intuition, note that Emma will always be less tempted to donate in the cold state than in the hot state. If she fears excessive donations in the hot state—an urge she can only control by being more ethical in round 1—she is more willing to violate a norm when the donation option is presented in a cold state. Hence, for a given transfer to others, if $p_G < 1$, Emma can achieve a higher own-consumption in Γ_c than in Γ_h. But because she also accumulates more guilt, she will subsequently donate a greater portion of her income. Unless further restrictions are imposed, the second effect can outweigh the first. It is always true however that if $\pi_c^*(a) < \pi_h^*(a)$, then Emma engages in a greater norm violation in Γ_c and gives more to others overall, $\pi_c^*(b) \geq \pi_h^*(b)$.

Proposition 2 has a simple corollary. If Emma knows in advance that she will be asked to donate in the future and donations are payoff inefficient, i.e., $p_G \geq 1$, such expectations deter norm violations in the present. The simplest way to describe the deterrence effect is to consider Emma's behavior in the absence of a future donation option.

Corollary 1 *Consider* $\Gamma_n = \{\Pi, \emptyset, \emptyset\}$ *and suppose* $p_G \geq 1$. *It follows that* $\pi_c^*(a) \leq \pi_n^*(a)$, *and if* $\pi_{c,1}^* \in N_\Pi$, *then* $\pi_{n,1}^* \in N_\Pi$.*Furthermore,* $\pi_{h,1}^*(a) \leq \pi_{c,1}^* \leq \pi_{n,1}^*$.

Norm Violations and Altruism Importantly, in our model, not *all* future donation will discourage present norm violations. Note that there is a complementary relationship between norm violations and prosocial actions. Suppose Emma's current options were to choose ($\$20, \$10$) by lying versus ($\$10$, $\$20$) without deception. Here adding the option to donate $\$2$ to others might encourage Emma to lie, *even if* when only $(20, 10)$ and $(10, 20)$ were implementable but both without deception, she would always choose $(20, 10)$.

"Paying for one's sins" will make a guilt-prone Emma feel better, and help balance the utility loss from violating a norm. Thus the donation option here cam *encourage* norm violations and the willingness to violate a norm might hing positively on Emma's ability to donate close to the violation. Instead, Corollary 1 establishes a wedge between the demand for altruism ex-ante and ex-post, which absent an ex-ante commitment device, serves as a *deterrent.*[8]

Our last result compares the case when a weakly inefficient donation option precedes a potential norm violation—Emma could "pay for her sins" in advance—to the case when the donation option is available later. We assume again that the timing of the donation option is initially known.

Proposition 3 *Consider* $\Gamma_{pre} = \{G, \Pi, \emptyset\}$ *and* $\Gamma_{post} = \{\emptyset, \Pi, G\}$. *If* $p_G \geq 1$ *it follows that* $\pi^*_{pre}(a) \geq \pi^*_{post}(a)$.

The above result shows that if donations are solicited prior to a norm violation, Emma will act more selfishly than if the donations are solicited after the potential norm violation. Intuitively, since Emma does not experience guilt in round 1, her preferences are weakly more selfish than in round 2 or round 3. Since donations do not improve efficiency, although the overall utility consequences of a given norm violation in both problems are the same, Emma's final own allocation is greater when she is not tempted by donations ex post.

[8]The combination of the above two effects implies a potentially *non-monotonic* relationship between the size of a donation opportunity and norm violations. Suppose Emma can implement either a particular norm violation, $\pi \in N$ or another efficient but ethical choice π'. Suppose that the subsequent donation set G is capped, and consider a gradual raising of this cap. It may well be the case that initially raising this cap will encourage norm violations. Above a certain threshold however, the effect is reversed and causes Emma to refrain from violating the norm. Similarly, there is a potential non-monotonic relationship with regard to the timing and a donation option. Here both a too early and a too late option can discourage a given norm violation. See also Corollary 3.

2.4 Implications

Reversal and Preference for Delay An emotional response in our model leads to a dynamic impulse control problem: a norm violation produces an altruistic urge that, given rational expectations, Emma would like to control prior to violating a norm. One way to limit such deviations is for Emma to constrain the amount of money she takes with her when attending a subsequent charity event.[9] Prior to violating a norm, Emma will value such a commitment if $p_G > 1$. An alternative manifestation of this demand for commitment concerns the preference for the timing of the action. Under Proposition 2, Emma initially will always prefer to *delay* a donation option. This way she can face the donation option in a later "colder" state than in an earlier "hoter" state.

Corollary 2 *Suppose* $p_G \geq 1$. *Initially, Emma will prefer* Γ_c *to* Γ_h.

When future donations are necessary to realize certain efficient payoff allocations the same prediction need not hold. Since round 3 preferences (U_3) could be *more* selfish than those in round 1 (U_1), Emma may only be able to implement certain efficient allocations if her most guilty self made the donation. Thus she might prefer an earlier to a later donation option and might only violate a norm if an early option is present. Even in this case, there may be a clear theoretical relation between Emma's ex-ante preference for the timing of the donation option and her actual donation behavior. We return to this in Section 3 where we show in Corollary 3 that in our experimental paradigm there can be preferences both for early and late donations, but those who do not want to donate will have a clear preference for their donation option to arrive later rather than sooner. [See page 16.]

The rich temporal pattern of preference reversal in our model also helps identify norm violations in observable behavior. Suppose payoffs and procedures—such as actions used to implement payoffs—are both observable. Individuals in our model exhibit not only a separable preference for certain procedures over others, but upon implementing a particular observable payoff allocation through a norm violation, they exhibit a temporal preference reversal in a systematic direction. Hence, procedures that violate Emma's norms can be identified in a dynamic context.

Donation Solicitation The model also speaks to the literature on the "demand" side of charity (e.g., Landry, Lange, List, Price, and Rupp 2006; Della Vigna, List, and Malmendier,

[9] In a similar spirit, Fudenberg and Levine (2006) predict that individuals with impulse control problems bring only "pocket money" to a nightclub.

2011), which examines the factors motivating individuals to give. If the conditions of Proposition 2 are satisfied, organizations aiming to maximize donation revenues, $\pi_2(b)$ above, should solicit contributions unannounced shortly after opportunities for potential norm violations occur. Our predictions are also consistent with the findings of Della Vigna, List, and Malmendier (2011) who show that when people are informed in advance that they will be asked to donate, a significant portion pre-commit not to donate. This decreased donations by 28% to 42% relative to a surprise solicitation. The authors attribute this to the fact that people don't like to say "no" face-to-face, or to not open the door when solicited as opposed to avoiding the solicitation in writing. Our model offers a parsimonious alternative: the announced donation opportunity has a negative option value to the extent that fluctuations in guilt may cause Emma to donate more in the future than what she finds optimal at the time of the pre-announcement. When the solicitation is a surprise, no such prior commitment is available.

Projection Bias People might well be aware that certain events will temporarily change their emotions, but may not fully appreciate the extent of such a change. Above we assumed rational expectations, but evidence points to a systematic misprediction here. Specifically, Loewenstein, O'Donoghue, and Rabin (2003) argue that people exaggerate the similarity between their actual state-dependent tastes and their future tastes. The presence of such projection bias alone could not generate our results – given the preference for commitment and delay – it often reinforces our rational expectations based mechanism of conscience accounting.

In the Appendix, we discuss how one can incorporate such projection bias into our setup. Prior to a norm violation, a biased Emma will underestimate how guilty she will feel ex post. Once experiencing guilt, she will underestimate how quickly this feeling will subside. Our main predictions are thus robust to the presence of such biased beliefs. In the case of full projection bias, the behavior described by Proposition 2 becomes equivalent to the behavior in Proposition 1, hence the result there holds for all $p_G \in (0, \infty)$. [10]

[10]Projection bias also leads to identifiable differences. For example, when donations are inefficient, there should be no donations under Proposition 2 but there will be under projection bias. A simple example illustrates this. After noticing a sharp nail on the street, a busy person may just walk by without picking it up. However, a little later she may feel so guilty as to turn back, remove the nail, and only after this costly detour, to continue on with her journey—a mistake that follows from underestimating how guilty she would feel later on.

3 A Deception Game

3.1 Procedure

To study conscience accounting empirically, we conducted a two-stage experiment. First, participants could lie to increase their profits at the expense of another participant. Second, after choosing whether to lie, we gave participants the option to donate to a charity.

We used a setup similar to Gneezy (2005). In this two-player deception game, one player, the Sender, has private information and the other, the Receiver, makes a choice based on a message conveyed by the Sender. The payoffs for both players depend on the choice the Receiver makes. This type of situation can be modeled using a cheap-talk setting. We constructed payoffs such that lying (sending a "wrong" misleading message) resulted in a higher payoff for the Sender.

In the instructions (see Appendix), we told participants that the experiment had two possible payment outcomes. Although the Receiver's choice would determine the outcome, only the Sender knew about the monetary outcomes of each option—the Receiver had no information regarding the alignment of incentives. Hence the Receiver's payoff expectations need not influence the Sender's behavior, which implies that the Sender's choice can be viewed as an individual decision problem that does not take strategic considerations into account.

After choosing the message—whether to lie or not—Senders were given the option to donate to a charitable foundation. We presented this option either directly after the message choice or with some delay, and Senders were either aware or not of the subsequent option when choosing the message.

We recruited 242 undergraduate students at the University of California, San Diego. The rules of the experiment were both read aloud and presented in written form to the participants. We informed them that neither Sender nor Receiver would ever know the identity of the player with whom they were matched. Participants in both roles knew that 1 out of 10 students assigned to the role of Sender would be randomly chosen to be paid, and we would match those individuals with Receivers in a different class.

Senders could choose from one of 10 possible messages to send the Receiver. Each message was in the form of "Choosing _ will earn you more money than any other number," with the blank corresponding to a number from 0 to 9. We told the Sender that if the Receiver chose a number that corresponded to the last digit of the Senders Personal Identification number

(PID), both players would be paid according to payment Option Y, and if the Receiver chose any other number, both players would be paid according to Option X. We informed Senders of the monetary consequences of both Option X and Y, and that the Receivers were not informed of this. We constructed the payments such that Option Y earned the Receiver more money than the Sender, and Option X earned the Sender more money than the Receiver. Hence, if the Sender expected the Receiver to follow her message, she had a monetary incentive to send one that did not correspond to the last digit of her PID—to lie—so the Receiver would choose the wrong number.[11]

Table I presents the payoffs we used in the experiment. We designed the Incentive, Incentive Delay and Informed Incentive treatments such that if the Receiver chose the wrong number, the Sender stood to earn $10 more and the Receiver $10 less than if the Receiver chose the correct number. In the No Incentive treatment the Sender had no monetary incentive to lie: both the Sender and Receiver stood to potentially earn $10 less if the Receiver chose the wrong number.

All four treatments offered Senders the option to donate $2 to the Make-A-Wish foundation after they had chosen what message to send. In the Incentive and No Incentive treatments, we presented the donation option directly after Senders made their message choices. In the Incentive Delay treatment, we presented the donation option with some delay: after their message choice, Senders received anagrams to solve for 10 minutes before we presented them with the option to donate. Importantly, in these three treatments Senders were not aware of the subsequent donation option when choosing what message to send, but were informed of it only after they made their initial choice.

In the Informed Incentive treatment, however, Senders knew in advance they would have the opportunity to donate. Particularly, we asked them to choose whether they wanted to make the decision to donate sooner (directly after their message choice) or later (at the end of the experiment), while at the same time deciding what message to send. Senders made the actual donation decision according to this choice. Ten minutes of anagrams once again served as the delay.

The last treatment was a baseline containing the same payoffs as the Incentive treatments

[11] In Gneezy (2005), the Sender could send one of two messages. Sutter (2009) showed that in a binary setting, a player who expects her partner to disbelieve her message may engage in sophisticated deception by sending a truthful message with the intention to deceive. To address these concerns, we used a message space with 10 possible messages. In our experiment, 75 percent of participants chose to follow the received message.

but excluding the donation option.

We established subject identification through the PID numbers the students provided as part of the experiment. We used the PID numbers to pay the participants according to the outcome of the experiment and to determine whether the Sender had lied in her message. Donations were \$2 in each available case, and we deducted the amount from the Senders' payments if they chose to donate. We then made the donations on the Senders' behalf directly through the Make-A-Wish website.

The set of message choices now corresponded to the allocation set Π, and the donation option to the donation set G. The logic of Proposition 1 is directly applicable. Senders who sent a false message in the Incentive treatment should be more likely to donate than those who sent a correct message, and these donation rates should be higher than for those who lied in the Incentive Delay treatment. Additionally, overall donation rates should be lower if the donation option was presented after some delay than directly after the message choice.

In the Informed Incentive treatment, Senders initially also made the choice of whether to be presented with the donation option sooner or later and only after this could they send a message. Given our binary setup, the theory here allows for an initial preferences in both direction, i.e. both for the hot and the cold decision environments. As discussed in section 2.4, relative to her round 1 preferences, U_1, Emma's preferences in the last round, U_3, may be too selfish. Therefore, Senders may choose to make their donation decisions earlier because they believe they will not donate if the option were presented later.[12]

Even in this binary setting however, the theory makes clear predictions on how the ability to choose the timing of the donation initially will affect actual *donation behavior*. Though the timing of the donation option will affect Emma's willingness to lie, it follows that in equilibrium her donation in the choice condition will be weakly greater than in the cold treatment and weakly lower than in the hot treatment. The subscript *choice* below refers to the case where a person selects into her preferred environment.

Corollary 3 *Let* Π *and* G *each be binary choice sets with payoff-undominated allocations such that again* $\underline{0} \in G$. *Then* $\pi^*_{c,2}(b) \leq \pi^*_{choice,2}(b) \leq \pi^*_{h,2}(b)$.

Since donations are efficient, in equilibrium the timing of the donations affects the incentives

[12]The source of preference heterogeneity in our model can be attributed to differences in γ (the speed at which guilt decays) or the curvature of u.

to lie: individuals may choose to violate a norm only when the donation option is available early or only when it is available late. Note first however, that if a person does not lie in either of the two conditions, her donation behavior is constant. If she only lies when the donation option is presented late, she will not donate in either of the treatments, and will *reveal* a preference for the late donation option. If she only lies if a donation option is presented early, then she will only donate when donation is early and will *reveal* a preference for the early donation option. Finally, in the case where she lies in both exogenous treatments, the comparision holds mechanically.

3.2 Results

Lying rates by treatment are presented in Table I. The differences in lying rates between the Incentive and Baseline treatments (Z=1.67, p=.10), Incentive and Incentive Delay treatments (Z=1.02, p=.15), and Incentive and Informed Incentive (Z=.43, p=.33) were not statistically significant.[13] However, differences between the Incentive and No Incentive treatments (Z=4.32; p<.001) and between the Baseline and No Incentive treatments (Z=5.79; p<.001) were statistically significant.

Our first key finding in this section is that in the Incentive treatment, when the donation option came as surprise directly after the message choice, 30% (6) of the participants who told the truth chose to donate, compared to 73% (27) of those who lied (Z=3.14; p<.001): the participants who chose to lie—and potentially earn $10 from lying—were significantly more likely to donate to charity than those who chose to tell the truth. This finding is not consistent with classifying individuals into simple "types" where some always behave in a moral way and others never do. In our experiment, those who donated to charity were *also* more likely to have previously lied.

However, in the Incentive Delay treatment, where the option to donate was presented some time after the message choice, 33% (3) of the participants who sent a true message chose to donate compared to 52% (14) of those who lied (Z=.96; p=.17). Particularly, those who lied and had the opportunity to donate directly after their message choice, did so significantly more often than those who lied and faced a delay between the two choices (Z=1.74; p=.04).

These results are summarized in Figure I. Particularly, they provide direct support for Propo-

[13] p-values were calculated from a one-tailed test of the equality of proportions using a normal approximation to the binomial distribution.

sition 1, which predicts that when the subsequent donation option comes as a surprise, individuals who violated a norm will be more likely to donate than those who did not, and that overall donations will be lower if the option is presented with some delay.[14]

Looking to the Informed Incentive treatment, we test the predictions of our model when the Sender knows about the donation option in advance. Here, 33% (5) of the Senders who told the truth chose to donate, compared to 57% (13) of those who lied—a weakly significant difference (Z=1.40; p=.08). In addition, of those who lied, 43% (10) of Senders chose to make their donation decisions early and 57% (13) chose to make their donation decisions late.

Of those who lied and chose to make their donation decisions early, 90% (9) actually donated, compared to 31% (4) of those who chose to decide later (Z=2.84, p<.001), as illustrated in Figure II. Furthermore, the overall donation rate in the Informed Incentive treatment (47%) was between that of the Incentive Delay (47%) and the Incentive (58%) treatments.[15]

To determine the extent to which these results represent conscience accounting rather than an income effect resulting from a higher expected payoff from deception, we compare the results of the Incentive treatment to those of the No Incentive treatment. In the No Incentive treatment, Senders did not have a monetary incentive to lie. Particularly, the Senders' expected payoff for lying in the Incentive treatment was the same as the expected payoff for truth in the No Incentive treatment. Here again the donation option was presented directly after message choice as a surprise. If differences in donation rates of liars and truth tellers had been due to an income effect, then those who lied in the Incentive treatment should have donated at the same rate as those who told the truth in the No Incentive treatment, since both choices had the same higher expected payoff of $20 rather than $10. However, the results do not support the income effect explanation. In the No Incentive treatment, of those who told the truth, 51% (21) chose to donate compared to 73% (27) of those who lied in the Incentive treatment. Those who lied in the Incentive treatment were still significantly more likely to donate than those who had told the truth in the No Incentive treatment (Z=1.97; p=.02), despite the fact that the expected own payoffs were the same.

[14]The difference in overall expected earnings of Senders was weakly significant (t=-1.29; p=.09).

[15]It should be noted that here we make the comparison when the donation option is unexpected, whereas in Corollary 2 the option is expected. Hence, to the extent that the extensive margin, i.e., the willingness to lie, is only moderately effected, our result serves as a good approximation.

4 An Over-paying Experiment

4.1 Procedure

In the deception game experiment, participants knew we were able to observe whether they lied. We designed the second experiment such that participants were unaware we were studying their moral choices. This unawareness should reduce behavior based on the experimenter demand effect and/or experimenter scrutiny.

We paid groups of subjects for their participation in an unrelated experiment. Two groups received payment according to how much we promised them. A third group received more than they were promised by "mistake" and had the opportunity to either return or keep the extra money.[16] We then gave all three groups the option to donate (not anticipated in advance) and recorded donation rates across the groups. In accordance with Proposition 1, we expected conscience accounting to manifest itself in the third group, predicting participants who decided to keep the extra money for themselves would be more likely to donate, and hence overall donation rates should be highest in the Mistake treatment.

We recruited 160 undergraduate students at the University of California, San Diego to participate in a coordination game experiment (see Blume and Gneezy, 2010). We invited subjects to the lab in pairs and seated them far apart for the duration of the game, which took approximately 15 minutes. We guaranteed all participants a $5 show-up fee, and those who did not succeed in coordinating did not get any extra money.

In addition, participants received $10 or $14, depending on the treatment, if they were able to coordinate with the individuals with whom they were matched. We randomly assigned those who had succeeded in coordinating to one of three treatments. In the Low treatment, we told subjects they would receive an additional $10 if they had succeeded in coordinating with their partners. In the High treatment, we told them the additional payment would be $14. In the Mistake treatment, we informed participants they would get $10 if they had succeeded, but we gave them $10 and an extra $4 by "mistake" : nine $1 bills and one $5 bill interspersed among them. Table II summarizes payments for all three treatments. After receiving their pay at the end of the experiment, participants in all three treatments received a description of a child with

[16]The study of individuals who do not know they are participating in an experiment is a common practice in field experiments, and is used in part to minimize experimenter demand effects that may be present in the lab.

cancer and were asked if they wanted to donate $1 from their final payment to the child.

When they received their pay, participants were told, "Here is your _. Please count it and sign this form," with the blank corresponding to the promised payment ($10 in the Low and Mistake treatments, $14 dollars in the High treatment). Then the experimenter left the room. All payments were made in $1 bills, except for the extra $5 bill in the Mistake treatment. Participants in all three treatments then decided whether to donate.

4.2 Results

In the Mistake treatment, 41% (33) participants returned the extra money they had received by "mistake." Donation rates by treatment are presented in Figure III. Overall, 30% (12) of participants in the Low, 25% (10) of those in the High and 49% (39) of those in the Mistake treatments donated. Consistent with conscience accounting, of those who returned the extra money in the Mistake treatment, 27% (9) made a donation, whereas 64% (30) of those who did not return the extra money made a donation ($Z=3.22$; $p<.001$). The overall donation rate in the Mistake treatment was significantly higher than in both the Low ($Z=1.96$; $p=.03$) and the High ($Z=2.50$; $p=.01$) treatments.

In addition, an income effect of earning $14 rather than $10 does not explain the discrepancy in donation rates. Subjects in the High treatment, who earned and were promised $14 before the experiment, donated at about the same rate as those who returned the extra money, but significantly less than those who kept it. Namely, although the donation rate for participants who returned the extra money is similar to those in the Low ($Z=.17$; $p=.43$) and High ($Z=.22$; $p=.41$) treatments, the donation rate for those who kept the money is significantly higher ($Z=3.15$; $p<.001$ and $Z=3.62$; $p<.001$, respectively). The difference in behavior in the Mistake treatment also suggests many participants, including those who did not return the money, did notice the mistake.

The results shown in Figure III also speaks to a "moral licensing" hypothesis proposed by Monin and Miller (2001), where past moral actions can justify less moral choices down the road. For example, the authors showed that participants allowed to establish themselves as not being prejudiced were more likely to later make remarks deemed socially offensive. One way to interpret moral licensing in the context of our experiment is to say that people who behaved morally and returned the extra money rather than achieved the same payoff without such a

moral act would be less likely to subsequently choose to donate because they had earned the "license" not to. Given this interpretation, the results presented in Figure III do not provide support for the moral licensing phenomenon. Consistent with our theoretical framework, people who returned the extra money, and hence did not violate a norm, donated at the same rate as those who had no option to make such a moral choice.

It should be noted that an important feature of studies demonstrating licensing is that the initial prosocial act was costless to the subject. For example, the subjects in the Monin and Miller (2001) study had the opportunity to establish themselves as unprejudiced at no cost to themselves. Khan and Dhar (2006) demonstrated licensing by having a group of individuals engage in one of two hypothetical volunteer assignments; they were then more likely than controls to choose a luxury item over a necessary item. However, a recent study by Gneezy, Imas, Nelson, Norton, and Brown (2011) found that cost is a critical factor in licensing, showing that when the initial prosocial act came at a cost to the subject, the licensing effect disappeared.

5 Discussion and Conclusion

In this paper, we formally examine emotional dynamics in the context of social behavior. We posit a theory where individuals care about the procedural aspects of their choices and, upon violating a norm, exhibit a specific time-inconsistency in their attitude towards others. This suggests an additional explanation for charitable behavior: people donate to account for their conscience after making a morally bad choice. The fact that people who lie are more likely to donate to charity than people who tell the (costly) truth may seem counter intuitive. One goal of this paper is to reshape this intuition.

Using experiments and a simple model, we show that in intertemporal choices the moral nature of a past choice impacts the nature of future choices in a systematic fashion. In our setup, past choices need to be "recent," but the definition of recent does not just depend on time. Simple other parameters that can go into the definition include the magnitude of the moral consequence of a choice or the bracketing rule that is used.

These findings are relevant in various economic situations. For example, travelers flying out of some airports receive the opportunity to offset the carbon footprint of their flight. Using "Climate Passport kiosks," people can calculate how many pounds of carbon dioxide their trip

will produce and the cost of offsetting this footprint using donations to programs aimed at greenhouse gas reduction. Several online travel retailers have begun to offer a similar option– giving customers the choice of offsetting their carbon footprint directly after ticket purchase. This kind of business is in line with the prediction of our model: people clear their bad feelings by donating. According to our model, programs that ask for donations close to the time of a purchase should be more successful than alternatives that ask people to donate at a remote (from a bracketing perspective) time.

Although the emotional response is temporary, it may be used strategically to increase prosocial acts or for organizations wishing to maximize donations. Furthermore, reminders of past unethical actions might lead to similar emotional dynamics as outlined in this paper. People may want to avoid being made to feel guilty, but nevertheless, will still act more prosocially if reminded about the ethical dimensions of past or current actions. If individuals are induced to feel guilty for having bought goods whose production has hurt others in an undue manner, they may have a greater propensity to opt for more expensive but fair products. Similarly, reminders of past immoral choices – such as broken promises or deceptions – can help organizations induce more loyalty or for charitable institutions to increase donations.

The results also highlight the importance of real temporal brackets in economic decisions. As mentioned before, the predictions are identified by the assumption that the emotional activation following an unethical choice is sufficiently fast. Indeed, evidence from neuroscience and psychology shows that the rise in emotional activation is typically much faster than the decline back to the neutral state (Garrett and Maddock, 2006). Although the implications of several major models of behavioral phenomena – e.g., Strotz (1955) – depend crucially on the specification of what the relevant time period is, there has been very little work on establishing the proper durations where the purported effects are the strongest. We believe that future studies connecting change in behavior and various measures of emotional activation in real time may provide key novel insights.

Throughout the paper, we have focused on the specific emotion of guilt. However, other negative *retrospective* emotions such as anger may fit a very similar temporal pattern in the context of social behavior (Card and Dahl, 2011). While guilt changes preferences to be more altruistic, events that provoke anger affect preferences so that hurting the other party becomes subsequently more desirable. Angry individuals may lash out at others even at a cost to them-

selves if such an opportunity arises soon after a trigger, but may prefer to control this impulse ex ante. In this manner, anger functions as a temporal shock to preferences directed against the payoff of others. Such effects of anger on decision making are greater immediately after the incitement than after some delay—consistent with the folk wisdom of anger management: "count to 10 before reacting."

Incorporating the emotional dynamics that lead to conscience accounting into models of charitable giving and prosocial behavior would provide further insight for theory that aims to better understand both the incidence of norm violations and altruism. Additionally, the general relationship between emotions and decision making outlined in our model provides an important avenue for future research, both on how emotions affect economic choices and the ways in which these effects are used strategically by individuals and organizations.

6 Appendix

Proof of Proposition 1. Since preferences are strictly convex, and the perceived problems in round 1 are identical, the initial choices in $\Gamma_{\widetilde{h}}$ and $\Gamma_{\widetilde{c}}$ are the same. Furthermore, if $d_2 = 0$, the continuation behaviors are also identical. Suppose now that $d_2 > 0$. Since moral debt is a bad and it shifts preferences in an altrusitic direction, as shown below, given norm monotonicity by continuity this optimum is unique in Π. Compare the marginal rates of substitutions in decision rounds 2 and 3. Given that $\gamma \in (0,1)$ it follows that for any π

$$\begin{aligned} MRS_h(\pi) &= \frac{u_{\pi_a}(\pi, d_2) + u_{\pi_a}(\pi, \gamma d_2)}{u_{\pi_b}(\pi, d_2) + u_{\pi_b}(\pi, \gamma d_2)} \leq \frac{u_{\pi_a}(\pi, \gamma d_2) + u_{\pi_a}(\pi, \gamma d_2)}{u_{\pi_b}(\pi, d_2) + u_{\pi_b}(\pi, \gamma d_2)} \\ &\leq \frac{u_{\pi_a}(\pi, \gamma d_2) + u_{\pi_a}(\pi, \gamma d_2)}{u_{\pi_b}(\pi, \gamma d_2) + u_{\pi_b}(\pi, \gamma d_2)} = MRS_c(\pi) \end{aligned}$$

Hence, given the necessary conditions for optimum, for any $p_G \in (0, \infty)$ the result follows. By totally differentiating the first-order condition in round 2, it follows that in the continuation strategy own-comsumption decreases in d_2 and $\frac{d\pi_a}{dd_2} = \frac{u_{\pi_b,d} - p_G u_{\pi_a,d}}{p_G u_{\pi_a,\pi_a} - 2u_{\pi_a,\pi_b} + \frac{1}{p_G} u_{\pi_b,\pi_b}} < 0$, equivalently the third round decision decreases in γ. ■

Proof of Proposition 2. Consider first the case where $p_G = 1$. Suppose $\pi_c^*(a) < \pi_h^*(a)$. Note first that Emma can implement π_h^* in Γ_c. Consider an initial choice of $\hat{\pi}_{c,1} = \pi_h^*$. Given the strict payoff concavity of u and the fact that $d_3(\hat{\pi}_{c,1}) \leq d_2(\hat{\pi}_{c,1}) \leq d_2(\pi_{h,1}^*)$ it follows from the optimality of π_h^* that $\hat{\pi}_c = \pi_h^*$. Similarly, in this case π_c^* is also implementable in Γ_h. Consider an initial choice $\overline{\pi}_{h,1} = \pi_c^*$. From the assumption that $\pi_h^*(a) > \pi_c^*(a)$, it follows that $\bar{\pi}_h = \pi_c^*$. Since round 1 preferences (U_1) are strict and identical, it must follow that $\pi_c^*(a) = \pi_h^*(a)$, a contradiction.

Consider now the case where $p_G > 1$. We show that if a final allocation π^* is optimal, then it can be constructed as a point on the frontier of Π. Consider Γ_h and suppose in contrast that $\pi_{h,2}^*(a) < 0$. Here one can always pick a final allocation π' such that $\pi'_{h,1}$ is on the payoff-efficiency frontier of Π and $\pi'_{h,1}(b) = \pi_{h,1}^*(b) + \pi_{h,2}^*(b)$. Furthermore, there always exists round 2 choice $\pi'_{h,2}$ such that $\pi'(a) > \pi^*(a)$. Let M denote Emma's continuation income after this initial choice– equal to her own payoff in the case where $\pi'_{h,2}(a) = 0$. Since the problem is separable

in time, it follows from strict quasi-concavity that holding d constant $\frac{d\pi_a}{dM} > 0$.[17] In addition, holding M constant, $\frac{d\pi_a}{dd_2} < 0$. Hence the following statements must be true: $\pi'(a) \geq \pi(a)^*$ and $\pi'(b) \geq \pi^*(b)$ and $d_2(\pi'_1) \leq d_2(\pi^*_1)$ with at least one of the inequalities holding strict. This however contradicts the optimality of π^*.

Given this fact, if there is a deviation in round 3 from π'_1 in Γ_c, then there is a deviation from π'_1 in round 2 in Γ_h. Since round 1 preferences are identical, it follows that $\pi^*_c(a) \geq \pi^*_h(a)$. ■

Proof of Corollary 1. Suppose in contrast that $\pi^*_{n,1}(a) < \pi^*_{c,1}(a)$. Let $\widehat{\pi}$ be the final allocation in Γ_c when $\widehat{\pi}_1 = \pi^*_n$ is combined with optimal continuation strategy thereafter. As long as $p_G \geq 1$ it follows that $\widehat{\pi}_3 = \underline{0}$, given that $d_3(\widehat{\pi}_1) \leq d_3(\pi^*_{c,1})$ and that $\pi^*_{c,3}(\pi^*_{c,1}) = \underline{0}$.[18] Thus π^*_n is implementable as a plan in Γ_c. Note however that π^*_n is maximal in Π given U_1. Hence a contradiction. ■

Proof of Proposition 3. By the logic of Proposition 2 an optimal allocation in Γ_{pre} can be constructed as a solution to $\Gamma_{\widehat{n}} = \{\emptyset, \Pi, \emptyset\}$. The result then follows from Corollary 1. ■

Proof of Corollary 2. The claim follows from the fact that in G only inefficient altruistic devations are possible and for any given π_1 Emma's continuation strategy is more altruistic in round 2 than in round 3. ■

Proof of Corollary 3. Consider first the case where $\pi^*_{c,1} = \pi^*_{h,1}$, then it must be true that $\pi^*_{c,3}(b) \leq \pi^*_{h,2}(b)$. Consider now the case where $\pi^*_{c,1}(a) > \pi^*_{h,1}(a)$, then if $\pi^*_{c,3}(b) > 0$, it follows that π^*_c is implementable in Γ_h. Also, by construction, π^*_h is implementable in Γ_c. Hence, they cannot generically be both be round 1 optimal and hence $\pi^*_{c,3}(b) \leq \pi^*_{h,2}(b)$. Finally in the case where $\pi^*_{c,1}(a) < \pi^*_{h,1}(a)$, if $\pi^*_{c,3}(b) > 0$, then π^*_c is implementable in Γ_h. If $\pi^*_{h,2}(b) = 0$, then π^*_h is implementable in Γ_c. Again, by the virtue of the same argument, they cannot differ and both be round 1 optimal, and hence $\pi^*_{c,3}(b) \leq \pi^*_{h,2}(b)$. ■

Projection Bias. We introduce projection bias to the specific setup of Section 2.4. Let us define an $\alpha-$biased,$\alpha \in [0,1]$, Emma's conditional expectations of d_t in period 1 given $\pi_1 \in \Pi$

[17] These follow from the facts that $\frac{d\pi_a}{dM} = \frac{u_{\pi_b,\pi_b} - p_D u_{\pi_a,\pi_b}}{p_D^2 u_{\pi_a,\pi_a} - 2p_D u_{\pi_a,\pi_b} + u_{\pi_b,\pi_b}} > 0$

[18] If $p_D = 1$ and $d_3(\widehat{\pi}_1) = 0$, it is without loss of generality to consider the implementation of a fixed final allocation with minimal donation in D.

to be

$$E_1^{\alpha}[d_t \mid \pi_1] := \alpha d_1 + (1-\alpha)d_t(\pi_1) = (1-\alpha)d_t(\pi_1), \tag{4}$$

for all t where $d_t(\pi_1)$ is the true d given π_1. For a given α, let $\pi_1^{*,\alpha}$ be the first element of a round 1 perceived optimal plan given α-biased expectations. It follows from Proposition 1 that $\pi_1^{*,\alpha}(a)$ is weakly increasing in α given norm-monotonicity and $u_d < 0$.

In the same way as above, let

$$E_2^{\alpha}[d_3 \mid d_2] = \alpha d_2 + (1-\alpha)d_3 = d_3 + \alpha(1-\gamma)d_2,$$

and $E_3^{\alpha}[d_3 \mid d_3] = d_3$. Let again be $\pi_2^{*,\alpha}(\pi_1)$ and $\pi_3^{*,\alpha}(\pi_1)$ be the optimal continuation strategies given a round 1 choice π_1 and α-biased expectations. It follows that for any given initial choice π_1, the difference in the $\alpha-$biased optimal continuation strategies $\pi_3^{*,\alpha}(a) - \pi_2^{*,\alpha}(a)$ is decreasing in α since $d_2 \geq 0$. Hence Proposition 1 extends.

In the case of expected donations, Emma will choose an allocation on the frontier of Π potentially incorrectly expecting not to deviate later on. Hence, believing that she will not later deviate, it follows from the proof of Proposition 2 that $\pi_{c,1}^{*,\alpha}(a) \geq \pi_{h,1}^{*,\alpha}(a)$. Also, since when $\alpha = 1$, $\pi_{c,1}^{*,\alpha}(a) = \pi_{h,1}^{*,\alpha}(a)$ here Proposition 2 holds for all p_G. ■

Table A.1

Table I: Results by Treatment

Treatment	**Option**	**Sender($)**	**Receiver($)**	***N***	**Lying(%)**	**Expected Earnings($)**
Incentive	X	20	10	57	65	15.3
	Y	10	20			
Incentive Delay	X	20	10	36	75	16.6
	Y	10	20			
Informed Incentive	X	20	10	38	61	15.1
	Y	10	20			
Baseline	X	20	10	57	79	17.9
	Y	10	20			
No Incentive	X	10	10	54	24	16.4
	Y	20	20			

Table A.2

Table II: Payoffs Used by Treatment

Treatment	**Payment Promised($)**	**Money Given by Mistake($)**	**Donation($)**	***N***
Low	10	-	1	40
High	14	-	1	40
Mistake	10	4	1	80

Figure A.1

Figure I: Fraction of Senders Who Donated by Message Type

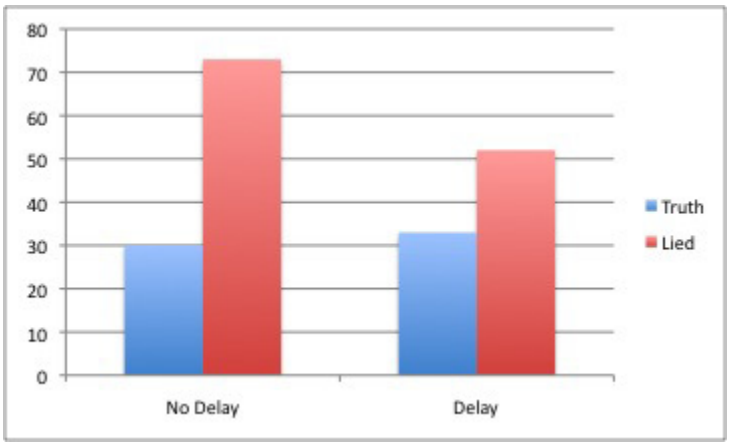


Figure A.2

Figure II: Fraction of Liars Who Donated by Timing of Decision

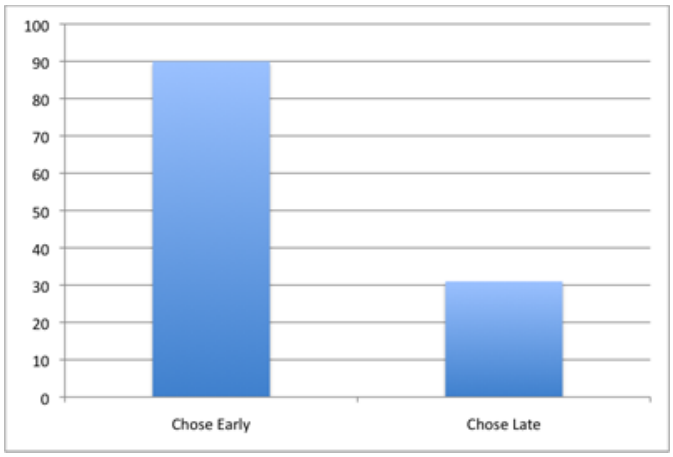
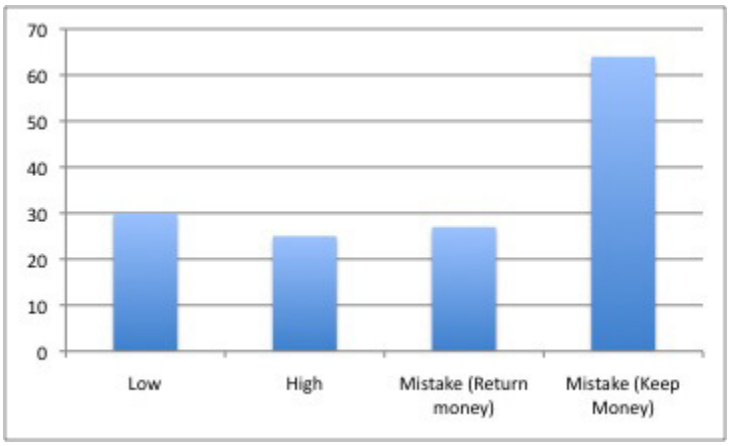


Figure A.3

Figure III: Fraction of Participants Who Donated by Treatment



References

[1] Akerlof, George and Kranton, Rachel E. "Economics and Identity." *Quarterly Journal of Economics*, (2000), 115, 715-53.

[2] Akerlof, George and Kranton, Rachel E. "Identity and the Economics of Organizations." *Journal of Economic Perspectives*, (2005), 19, 9-32.

[3] Andreoni, James. "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving." *Economic Journal*, (1990), 100, 464-77.

[4] Andreoni, James. "Warm-Glow Versus Cold-Prickle: The Effects of Positive and Negative Framing on Cooperation in Experiments." *Quarterly Journal of Economics*, (1995), 110, 1-21.

[5] Battigalli, Pierpaolo and Dufwenberg, Martin. "Guilt in Games." *American Economic Review*, (2007), 97, 170-76.

[6] Battigalli, Pierpaolo and Dufwenberg, Martin. "Dynamic Psychological Games." *Journal of Economic Theory,* 2009, 144, 1-35.

[7] Baumeister, Roy F.; Stillwell, Arlene M., and Heatherton, Todd F. "Guilt: An Interpersonal Approach." *Psychological Bulletin*, (1994), 115, 243-67.

[8] Becker, Gary S. "Altruism, Egoism, and Genetic Fitness: Economics and Sociobiology." *Journal of Economic Literature*, (1976), 14, 817-26.

[9] Blume, Andreas and Gneezy, Uri. "Cognitive Forward Induction and Coordination without Common Knowledge: An Experimental Study." *Games and Economic Behavior*, (2010), 68, 488-511.

[10] Card, David and Gordon Dahl. "Family Violence and Football: The Effect of Unexpected Emotional Cues on Violent Behavior." *Quarterly Journal of Economics*, (2011), 126, 1–41.

[11] Charness, Gary and Dufwenberg, Martin. "Promises and Partnerships." *Econometrica*, (2006), 74, 1579-1601.

[12] Charness, Gary and Rabin, Matthew. "Understanding Social Preferences With simple tests." *Quarterly Journal of Economics*, (2002), 117, 817-69.

[13] Dana, Jason; Weber, Roberto A. and, Kuang, Jason X. "Exploiting Moral Wriggle Room: Experiments Demonstrating an Illusory Preference for Fairness." *Economic Theory*, (2007), 33, 67-80.

[14] Della Vigna, Stefano; List, John A. and, Malmendier, Ulrike. "Testing for Altruism and Social Pressure in Charitable Giving." *Quarterly Journal of Economics*, forthcoming.

[15] Dreber, Anna and Johannesson, Magnus. "Gender Differences in Deception." *Economic Letters*, (2008), 99, 197-99.

[16] Dufwenberg, Martin and Gneezy, Uri. "Measuring Beliefs in an Experimental Lost Wallet Game." *Games and Economic Behavior*, (2000), 30, 163-82.

[17] Elster, Jon. "Emotions and Economic Theory." *Journal of Economic Literature*, (1998), 36, 47-74.

[18] Fudenberg, Drew and Levine, David K. "A Dual Self Model of Impulse Control." *American Economic Review,* (2006), 96, 1449-76.

[19] Garrett, Amy S., and Maddock, Richard J. "Separating Subjective Emotion From the Perception of Emotion-Inducing Stimuli: An fMRI Study." *NeuroImage,* (2006), 33, 263-74.

[20] Gneezy, Ayelet; Imas, Alex O.; Nelson, Leif D.; Norton, Michael I., and Brown, Amber. "Paying to be Nice: Costly Prosocial Behavior and Consistency." *Management Science*, forthcoming.

[21] Gneezy, Uri. "Deception: The Role of Consequences." *American Economic Review*, (2005), 95, 384-95.

[22] Harris, Christopher. "Existence and Characterization of Perfect Equilibrium in Games of Perfect Information." *Econometrica*, (1985), 53, 613-28.

[23] Kahneman, Daniel; Knetsch, Jack L., and Thaler, Richard H. "Fairness and the Assumptions of Economics." *The Journal of Business*, (1986), 59, S285-300.

[24] Khan, Uzma and Dhar, Ravi. "Licensing Effect in Consumer Choice." *Journal of Marketing Research,* (2006), 43, 259-266.

[25] Kőszegi, Botond, and Rabin, Matthew. "Reference-Dependent Risk Attitudes." *American Economic Review,* (2007), 97, 1147-73.

[26] Landry, Craig; Lange, Andreas; List, John A.; Price, Michael K., and Rupp, Nicholas."Toward an Understanding of the Economics of Charity: Evidence from a Field Experiment." *Quarterly Journal of Economics*, (2006), 121, 747-82.

[27] Lazear, Edward P.; Malmendier, Ulrike, and Weber, Roberto A. "Sorting in Experiments with Applications to Social Preferences." *American Economic Journal: Applied Economics*, forthcoming.

[28] Loewenstein, George. "Anticipation and the Value of Delayed Consumption." *Economic Journal*, (1987), 97, 666-84.

[29] Loewenstein, George; O'Donoghue, Ted, and Rabin, Matthew. "Projection Bias in Predicting Future Utility." *Quarterly Journal of Economics,* (2003), 118, 1209-48.

[30] Meier, Stephan. "Do Subsidies Increase Charitable Giving in the Long Run? Matching Donations in a Field Experiment." *Journal of the European Economic Association*, (2007), 5, 1203-22.

[31] Monin, Benoit and Miller, Dale T. "Moral Credentials and the Expression of Prejudice." *Journal of Personality and Social Psychology*, (2001), 81, 33-43.

[32] Rabin, Matthew. "Moral Preferences, Moral Constraints, and Self-Serving Biases." *mimeo*, (1995), UC Berkeley.

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

[34] Sutter, Matthias. "Deception Through Telling the Truth?! Experimental Evidence From Individuals and Teams." *Economic Journal*, (2009), 119, 47-60.

[35] Vesterlund, Lise. "The Informational Value of Sequential Fundraising"*Journal of Public Economics*, (2003), 87, 627-57.