Moral licenses: Strong experimental evidence

Pablo Brañas-Garza, GLOBE: Universidad de Granada, Spain

Marisa Bucheli, Universidad de la República, Uruguay

Teresa García-Muñoz, GLOBE: Universidad de Granada, Spain

ABSTRACT- Research on moral cleansing and moral self-licensing provides a framework to explain the dynamics of moral behavior. Bad deeds trigger negative feelings that make people more likely to engage in moral behavior to offset them. Good deeds favor a positive self-perception that creates licensing effects that make people to engage in behaviors that are less likely to be moral. In short, a deviation from a "normal state of being" is balanced with a subsequent action that compensates the prior behavior. This paper reports experimental evidence that give support to the idea that actions are affected by past actions. To explore this phenomenon we run an economic experiment where subjects play a sequence of giving decisions (dictator games). The amount of money he/she kept in every round in analyzed using an estimation technique to accurately measure the dynamics of these actions. We find that past donations (only the previous one) affect actual decisions but the sign is negative: subjects change in every round what they did in the past (generous → selfish → generous). Hence donations over time are the result of a systematic process of equalization: moral licensing (being selfish after altruist) or cleansing (altruistic after selfish).

I. INTRODUCTION

How and *why* moral behavior emerges is a critical question. Being nice is not costless: every single altruistic action generates a cost for the donor. Thus, each good deed needs a benefit to be triggered. Despite a number of classical evolutionary arguments such us kin selection –Hamilton rule- or reciprocal altruism (Fehr and Fischbauer, 2003) another series of papers deal with more self-centered arguments like identity, guilt-aversion or warm-glow, that describe the benefits of being moral (see Akerlof and Kranton, 2000, Charness and Dufwenberg, 2006, Battigalli and Dufwenberg, 2007 and Aguiar et al. 2010). In this paper we are interested on the moral self-licensing and moral cleansing literature that proposes a framework in which past moral behavior affects the costs of being moral in the present.

One motivation of good deeds is their positive effect on moral self-worth. But when past actions make people to feel confident about their moral behavior, their moral self-regard could be high enough to allow them engaging in morally dubious behaviors in the present (Zhong and Liljenquist, 2006; Merritt, Effron and Monin, 2010). This is the central argument of moral self-licensing literature. In a review of the evidence, Merritt *et al* (2010) present the two more frequent moral-licensing mechanisms used in the literature: the moral credits and the credentials models. The moral credits model uses a moral bank account metaphor: good deeds purchase "moral credits" that diminish the uncomfort of engaging in bad deeds in the future. The credentials model states that prior actions affect the meaning of present actions. That is, an action in the past allows to interpreting the present action of a licensed person without morally ambiguity, in a context in which the same action may be a transgression for an unlicensed person. Note that in the first model, the licensed person involves in what he considers a bad action; in the second model, the action does not hurt his self-image. Thus, we may expect a less number of transgressions of a licensed person under the moral credits than the credentials mechanism.

In turn, immoral behavior has a negative effect on moral self-worth. After doing bad deeds, people engage in moral behavior to recover self-worth; this mechanism is the so-called moral cleansing behavior (see Sachdeva, Iliev and Medin, 2009). One well documented example is that in response to sins, many religious practices require

bodily purification.

Taking into account the two types of behavior, Sachdeva *et al*. (2009) consider "*moral behavior as being embedded within a larger system that contains competing forces. Moral or immoral action may emerge from an attempt to find balance among these forces*". There is a symmetric process: every deviation from the normal state of behavior is balanced with a subsequent more moral action (moral cleansing) or less moral action (moral licensing). Sachdeva *et al*. (2009) show that after a positive (negative) priming subjects are less (more) prone to make donations to charities.

This paper provides sound evidence of this phenomenon. We analyze data from an economic experiment where subjects played a sequence of 16 dictator games with distinct recipients (sharing a pie under anonymity conditions). In every decision we capture the percentage of the money that the donator keeps for himself. All the games are identical in the format but framed. Besides a blind (baseline) game, we use three types of frames regarding the information given about gender (male/female), income (poor/rich) and political preferences (right/left) to generate 15 different environments. Each subject received the 16 instructions consecutively and in different random order.

Using an estimation technique to accurately measure the dynamics of these actions we estimate how a donation (d_{t-1}) affects the subsequent one (d_t). We find that donations over time follow an AR(1)[1] process with negative coefficient that allows us stating two important conclusions:

i. the negative sign of immediate past actions (d_{t-1}) on current choices (d_t) indicates that in every round subjects revert what they did in the past;

ii. the length of the AR(1) indicates that only most immediate previous period affects present behavior. Hence, subjects tend to balance today what they did yesterday. This is not a long memory process.

The rest of the paper is organized as follows. Section 2 explains the experiment design and procedures. Third section shows the results and the fourth concludes. The econometric method is present in the appendix.

[1] AR(p) is an auto-regressive process of length p, bein p the number periods which affect actual values.

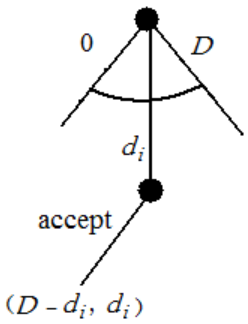
II. THE EXPERIMENT

The dictator game

In the dictator game (Forsythe et al., 1994), the first player, "the proposer", determines an allocation (split) of some endowment (such as a cash prize). The second player, "the responder", simply receives the share of the pie left by the proposer. The responder's role is entirely passive.[2]

Formally, given a pie of size D, the dictator must decide any value of $d_i \in [0, D]$ to pass to the recipient. Therefore the final distribution of benefits is as follows:

$$D\text{-}d_i \, , \, d_i$$



where $D\text{-}d_i$ is the dictator's benefit. Since the Nash equilibrium is giving zero, any strictly positive donation, $d_i > 0$, is interpreted as pure altruism.

Participants

176 subjects participated in the experiment (dictators and recipients). We will focus only in the sample of 88 dictators (32 % of women) since recipients do not play any role in our analysis. The participants were undergraduate students of several degrees at the *Universidad de la República* (Uruguay). All of them were volunteers who answered to a public call.

Procedures and materials

The experiment consisted of four sessions where subjects were given verbal and printed information: they had to take 16 decisions and each one was explained in one sheet of a printed booklet. They were not allowed to speak to one another and they were seated in such a way that they could not see the written responses of the other subjects.

The baseline treatment consisted of a standard dictator game in which each participant was a dictator or a recipient (the participants knew that no one will play both roles).

---

[2] As a result, the dictator game is not formally a game at all. To be a game, every player's outcome must depend on the actions of at least some others. Since the proposer's outcome depends only on his own actions, this situation is one of decision theory and not game theory.

The dictator had to allocate 10 bills of 20 Uruguayan pesos (around 10 American dollars) between himself and a randomly chosen student placed at other room. Following List (2007) instructions, the task was explained in one sheet of a printed booklet and the possible payoffs were presented on a line in which the subject had to mark their decision with a circle. The amount of money ranked from 0 pesos (left-end) to 200 pesos (right-end) and the donations were restricted to multiples of 20 including zero.

The rest of the treatments were identical to the baseline (blind) with the exception of the framing. In order to frame the task we used information that participants gave at the moment of the inscription to the experiment: sex, income category and ideological category. With this information we labeled the participants as women/men, rich/poor and right-wind/left-wind.[3]

In three treatments, the donator was told that the recipient will know the donator's sex, income category or ideological category, respectively. In six treatments, the donator knew one characteristic of the recipient (sex or income category or ideological category). In another six treatments, besides knowing one characteristic of the recipient, the donator was told that the recipient will know the game's framing (for example, the recipient will know that the donation was done from a woman to a man).

The entire booklet consisted of sixteen tasks that were presented in a different random order for each subject. The different random order is an important characteristic of the design: as in each round the donators are facing different frames, there is no room for an equalizing pattern in the course of the experiment common to all subjects.

We paid only one decision (randomly chosen) to each dictator which avoids the effect of accumulation of earnings in the course of the session. Besides, the use of different recipients and frames in each decision helped to maintain subjects' interest. Notice that once a decision is paid, the subsequent decisions cannot actually hurt or help the recipient. Thus, if the donor takes what he thinks a selfish (generous) decision, the

[3] We asked the participants to fill a questionnaire where they informed the socio-economic position of their household and their ideological position in a 10-steps scale where 1 was extreme poor/left and 10 was extreme rich/right. In order to build binary labels (poor/rich, left-wind/right-wind), the separation threshold was the median value of the reported distributions.

subsequent action will not compensate the prior recipient because of two reasons: only one decision is paid and the recipients are different subjects.

The money transferred to recipients was given to them in a different session where they were cited. Considering all the games, the average dictators earning was U$142.5 (7 bills) and, consequently, the mean recipient earning was U$57.5 (3 bills).

III. RESULTS

We use a dynamic panel data model to estimate the donation at period *t* where the immediate past donation (d_{t-1}) is an explanatory variable of current decision:

$$d_{it} = \alpha_i + \gamma d_{i,t-1} + x_{it}^{'}\beta + v_{it}, \quad i = 1,...,88 \text{ individuals}, \quad t = 1,...,16 \text{ rounds}$$

where α_i denotes an unobserved individual-specific time-invariant effects, in our several models the regressors x_{it} will be treatment dummies and temporal trend and they are uncorrelated with $v_{it}\ \forall i,t$ (strictly exogenous regressors), the disturbance terms v_{it} are independent and identically distributed $\forall i,t$. The individual effects can be of fixed or random nature. It not easy to choose between both model but random effects model is appropriated when individuals are chosen randomly from a large population. On the contrary, the fixed effect model is appropriated when analysis is focused in a specific set of individuals, as it is our case. However, with random effects model we have obtained similar results.

We use two-step GMM estimators with the Windmeijer correction using lagged levels (*t-2*, *t-3* and *t-4*) of the dependent variable as instruments[4] (Arellano and Bond, 1991; Windmeijer, 2005; see appendix for a description of the technique).

Table 1 shows the results of three models. In Model 1, the only covariate is the immediate past donation (d_{t-1}); in Model 2 we also include the treatment dummies and in Model 3 we add a temporal trend. In the three estimations, the coefficient of past donation (d_{t-1}) is negative, significant and less than one in absolute value. Besides, the trend is no significant. In the bottom part of Table 1 we show Arellano-Bond tests to validate the instruments (see appendix). No any single test is rejected.

[4] For this reason we lost the observations of the first two rounds.

Table 1: All rounds

	Model 1	Model 2	Model 3
Round (t)	-	-	0.205
			(0.409)
d_{t-1}	**-0.089**	**-0.088**	**-0.074**
	(0.028)	**(0.032)**	**(0.033)**
Constant	**64.464**	**62.736**	**59.546**
	(0.000)	**(0.000)**	**(0.000)**
Treatment controls	not	yes	yes
Arellano-Bond serial	-0.665	-0.640	-0.504
correlation test	(0.506)	(0.522)	(0.614)
Instruments	*40*	*42*	*43*
Sample Size	*1217*	*1217*	*1217*

p-values in parentheses.

The important result here is that time series of donations follow an stationary AR(1) process with negative coefficient. This means that in successive periods, donations move around its mean but with a very noisy behavior, crossing constantly the mean. Hence, subjects tend to balance in a round what they did in the prior round.

Thus, we do not find a result favoring the consistency of preferences but an equalization behavior. The pattern of donations over time emerges as the result of a systematic process of equalization: moral licensing (being selfish after altruist) or cleansing (altruistic after selfish).

We also check if donations follow an AR(2) process: we find that the coefficients of d_{t-2} were never significant whereas the coefficients of d_{t-1} were still negative and significant (not reported, available upon request).

IV. ROBUSTNESS

As a simple robustness test, we check how our results change (or not) when we use different sample sizes. Table 2 shows the same models using the last 12 periods (*t=5, 6, ... , 16*) and the last 8 periods (*t=9, 10, ... , 16*). Recall that given that every individual played a different random order we miss different treatment's observations for each individual.

Table 2: Robustness

	Rounds 5 to 16			Rounds 9 to 16		
	Model 4	Model 5	Model 6	Model 4	Model 5	Model 6
Round (t)	-	-	0.088	-	-	0.161
			(0.714)			(0.727)
d_{t-1}	**-0.101**	**-0.097**	**-0.097**	**-0.120**	**-0.123**	**-0.129**
	(0.053)	**(0.066)**	**(0.040)**	**(0.075)**	**(0.060)**	**(0.015)**
Dummy 1	-	0.906	0.919	-	-0.139	-0.050
		(0.682)	(0.677)		(0.968)	(0.988)
Dummy 2	-	**4.086**	**3.989**	-	**6.344**	**6.034**
		(0.003)	**(0.004)**		**(0.047)**	**(0.023)**
Constant	**66.644**	**64.775**	**63.369**	**68.726**	**66.718**	**65.155**
	(0.000)	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**
Arellano-Bond serial correlation test	-0.614	-0.552	-0.577	-0.378	-0.350	-0.406
	(0.539)	(0.581)	(0.567)	(0.705)	(0.726)	(0.685)
Instruments	37	39	40	25	27	28
Sample Size	1043	1043	1043	695	695	695

p-values in parentheses.

The main message is that we do not observe remarkable differences when we compare results from table 1 and 2. Hence, using initial or final rounds of the experimental session does not provide any difference.

Lastly Table 3 shows a new exercise. We estimate the AR(*p*) model -with controls- for a sample of 68 subjects randomly selected, that is, we drop 20 subjects. We repeat the exercise removing another 20 different subjects and finally we repeat the removal a third time. Table 3 shows the estimated AR(1) coefficients for the three sub-samples (elimination #1, #2 and #3).

Table 3: Additional robustness tests

	AR(1) Coefficient	**p-value**	**Sample Size**
Renoval of 20 participants			
elimination #1	-0.099	0.033	944
elimination #2	-0.091	0.063	941
elimination #3	-0.077	0.069	940
Other trials			
without "Blind"	-0.122	0.017	989
without "Constant"	-0.075	0.032	1094

The last trials are shown in the bottom part of Table 3. We estimate the same for the resulting sample when observations from the baseline are not included. Results are even stronger (*p-value=0.01*) than previous ones. Also we run a model without constant where results are identical to those obtained previously.

So, we may conclude that subjects tend to balance today what they did yesterday.

V. CONCLUSIONS

This research joins the literature that focuses on the role of moral cleansing and moral self-licensing on behavior. Our results show that that over time, in a dictatorial game setting, the donations do not have a trend. However this stability across time cannot be interpreted as the result of strong preferences for altruism. In contrast, this stability emerges as the result of equalization. In the estimations, the past donation (d_{t-1}) coefficient is always negative, significant and its absolute value is less than one-indicating that subjects who behaved nicely yesterday are selfish today and vice versa. In short, a systematic moral self-licensing and moral cleansing pattern emerges.

There is an important consequence of this type of behavior for "current" theories of *Guilt* (Battigalli and Dufwenberg, 2007) and *Identity* (Akerlof and Kranton, 2000). Not only "bad" deviations from the normal state of behavior are balanced with a subsequent more moral action (moral cleansing) but also moral licensing is used after some periods of goodness. The later is not explained from these approaches.

References

Aguiar, F., Brañas-Garza, P., Espinosa, M. P. and L. Miller (2010). Personal identity. A theoretical and experimental analysis, *Journal of Economic Methodology* **17**(3): 261-275.

Akerlof, G. and R.E. Kranton (2000). Economics and Identity, *Quarterly Journal of Economics* **115**: 715–753.

Anderson, T.W. and C. Hsiao (1982). Formulation and estimation of dynamic models using panel data, *Journal of Econometrics* **18**: 47-82.

Arellano, M. and S. Bond (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations, *Review of Economic Studies* **58**: 277-297.

Battigalli, P. and M. Dufwenberg (2007). Guilt in games, *American Economic Review* **97**(2): 170-176.

Charness, G. and M. Dufwenberg (2006). Promises & Partnership, *Econometrica* **74**: 1579 -1601.

Fehr, E. and U. Fischbauer (1999). The nature of human altruism, *Nature* **425**: 1579 - 1601.

Forsythe, R., Horowitz, J.L., Savin N.E. and M. Sefton (1994). Fairness in Simple Bargaining Experiments, *Games and Economic Behavior* **6**: 347-369.

Hansen, L. (1982). Large sample properties of generalized method of moments estimation, *Econometrica* **50**(3): 1029-1054.

List, J. A. (2007). On the Interpretation of Giving in Dictator Games, *Journal of Political Economy* **115**(3): 482-93.

Merritt, A. C., Effron, D. and B. Monin (2010). Moral Self-Licensing: When Being Good Frees Us to Be Bad, *Social and Personality Psychology Compass* **4/5**: 344–357.

Sachdeva, S. Iliev, R. and D.L. Medin (2009). Sinning Saints and Saintly Sinners. The Paradox of Moral Self-Regulation, *Psychological Science* **20**: 523-528.

Windmeijer, F. (2005). A finite sample correction for the variance of linear efficient two-step GMM estimators, *Journal of Econometrics* **126**: 25-51.

Zhong, C. and K. Liljenquist (2006). Washing Away Your Sins: Threatened Morality and Physical Cleansing, *Science* **313**(5792): 1451-1452.

Appendix: Dynamic Panel Data

Panel data regressions allow us to study individual behavior in a repetitive environment. In turn, the use of dynamic panel data model allows the possibility of considering lags (donations done in previous round in our case). Arellano-Bond (1991) dynamic panel estimator is designed for a situation where the dependent variable depends on its own past realizations. The model is

$$d_{it} = \gamma d_{i,t-1} + x_{it}^{'}\beta + \alpha_i + v_{it}, \quad i = 1,...,88 \text{ individuals}, \quad t = 1,...,16 \text{ rounds}$$

where α_i captures unobservable heterogeneity and can be of fixed or random nature, the disturbance terms v_{it} are independent and identically distributed $\forall i,t$ and the regressors x_{it} are uncorrelated with v_{it} $\forall i,t$ (strictly exogenous regressors).

The dynamic structure of this equation make not valid the usual estimators in panel data (within estimators, generalized least squared estimators,…). Observe that, by construction, $d_{i,t-1}$ is correlated with the unobserved individual effect α_i. The first-differencing transformation removes the individual effect from the model:

$$\nabla d_{it} = \gamma \nabla d_{i,t-1} + \nabla x_{it}^{'}\beta + \nabla v_{it}$$

In this model, $\nabla d_{i,t-1}$ is correlated with ∇v_{it} (by construction). Anderson and Hsiao (1982) proposed a Two Stage Least Squares estimator using lagged levels of d_{it} as instruments for $\nabla d_{i,t-1}$. But, when the panel has more than three periods, additional instruments are available and the model is overidentified (it has more instruments than parameters). Arellano and Bond (1991) used Generalized Method of Moments (GMM; Hansen, 1982) to obtain parameter estimators by moment conditions generated by lagged level of the dependent variable with first differences of the error v_{it}.[5]

In the case of non-identically distributed disturbances, a two-step GMM estimator is used. The two-step GMM estimator is efficient and robust under heteroskedasticity but its standard errors have downward bias. To solve this problema, Windmeijer

[5] They also used first-differences of strictly exogeneous regressors to create moment conditions.

(2005) proposed a correction for the two-step standard errors. In the estimations of the models presented in the section III we use two-step GMM estimators with the Windmeijer correction.

Arellano and Bond (1991) developed a test to detect serial correlation in the first-differenced disturbances. This situation would do some lags invalid as instruments: if the v_{it} are serially correlated of order 1 then, $y_{i,t-2}$ is endogenous to ∇v_{it} (by the presence of $v_{i,t-1}$ in the difference) and $y_{i,t-2}$ would be an invalid instrument. If the null hypothesis of this test (there is not serial correlation) is not rejected, the validation of the instrumental variables is confirmed. In tables 1 and 2, we present the statistics and their associated p-values of this test.