Economics Letters 214 (2022) 110436

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

# Self-regulation after temptation?

Matthieu Légeret *, Christian Zehnder

*Faculty of Business and Economics, University of Lausanne, 1015 Lausanne, Switzerland*

ARTICLE INFO

*Article history:*
Received 21 December 2021
Received in revised form 1 March 2022
Accepted 9 March 2022
Available online 15 March 2022

*Keywords:*
Moral cleansing
Self-regulation
Behavioral experiment

ABSTRACT

Although moral cleansing—a form of self-regulation—has frequently been studied, existing evidence is mixed and its prerequisites remain unclear. We hypothesize that large, salient deviations from self-defined morality require regulation through moral cleansing, whereas small, inconspicuous deviations are tolerated and lead to continued misbehavior. Using an incentivized online experiment, we measure participants' baseline morality before using temptations to induce deviations. We find that weak temptations lead to small reductions in moral behavior that remain uncorrected. However, we observe that larger deviations induced by strong temptation do not lead to compensation.



## 1. Introduction

Moral choices sometimes exhibit dynamic patterns that seem paradoxical. For example, morally questionable actions may be followed by particularly moral behavior. Such patterns may reflect attempts to self-regulate one's moral identity: individuals whose past behavior conflicts with their own moral code may experience disgust (e.g., Chapman and Anderson, 2013) and may respond with "moral cleansing" (Lee and Schwarz, 2021; Gneezy et al., 2014; Sachdeva et al., 2009).

Although evidence for morally inconsistent behavior exists (e.g., Effron et al., 2009; Merritt et al., 2010; Zhong and Liljenquist, 2006), investigations of the underlying mechanisms yielded mixed results: some situations trigger moral cleansing while others do not (Dolan and Galizzi, 2015). Moreover, evidence also indicates that morality may erode over time, in that immoral behaviors are sometimes followed by even more immoral behaviors (Fischbacher and Föllmi-Heusi, 2013). Our paper introduces two elements that shed new light on the debate about moral cleansing.

First, at the conceptual level, we hypothesize that self-regulation is triggered by the saliency of misbehavior. We expect that large and salient deviations from self-defined morality require moral cleansing, whereas small and inconspicuous deviations are tolerated and lead to continued misbehavior. Our second contribution is methodological. In previous research baseline morality is typically not measured. The lack of a reference point for "normal" behavior makes it hard to establish whether actions identified as self-regulating behaviors really are exceptionally moral. In our incentivized experiment, we first measure individual initial morality before exposing participants to a stimulus designed to induce immoral behavior. In the final phase, we observe participants' reaction once the stimulus has been removed. Our treatments vary the intensity of the stimulus to trigger deviations of different magnitude and saliency.

## 2. Experimental design and procedures

We recruited 615 participants on Amazon Mechanical Turk, for an experiment with three parts, each consisting of 10 rounds. Participants' task was to count the ones in tables containing random sequences of ones and zeros (Abeler et al., 2011). In each round participants could choose between a large and a small table and a part was completed once 10 tables had been correctly completed (irrespective of the type of table). Each correctly completed table had an externality on a charitable donation: a large table increased the donation by 10 cents, whereas a small table removed 1 cent. Small tables therefore allowed participants to finish faster for the same wage, at the expense of the charity.

In the first part all participants faced the same conditions: small tables contained 6 lines, large tables contained 9 lines. In the second part, participants were randomly assigned to one of three experimental treatments. In the control condition (C) the size of small and large tables remained as in part 1. In the low (L) and high (H) temptation treatments, the size of the small tables was reduced to 5 and 3 lines, respectively. In the third part, the size of small tables was reverted to 6 lines for all participants.

## 3. Results

### 3.1. General pattern

The top panel of Fig. 1 shows the proportion of large tables among all completed tables in all three parts. Column (1) of

* Corresponding author.
*E-mail address:* matthieu.legeret@unil.ch (M. Légeret).

https://doi.org/10.1016/j.econlet.2022.110436

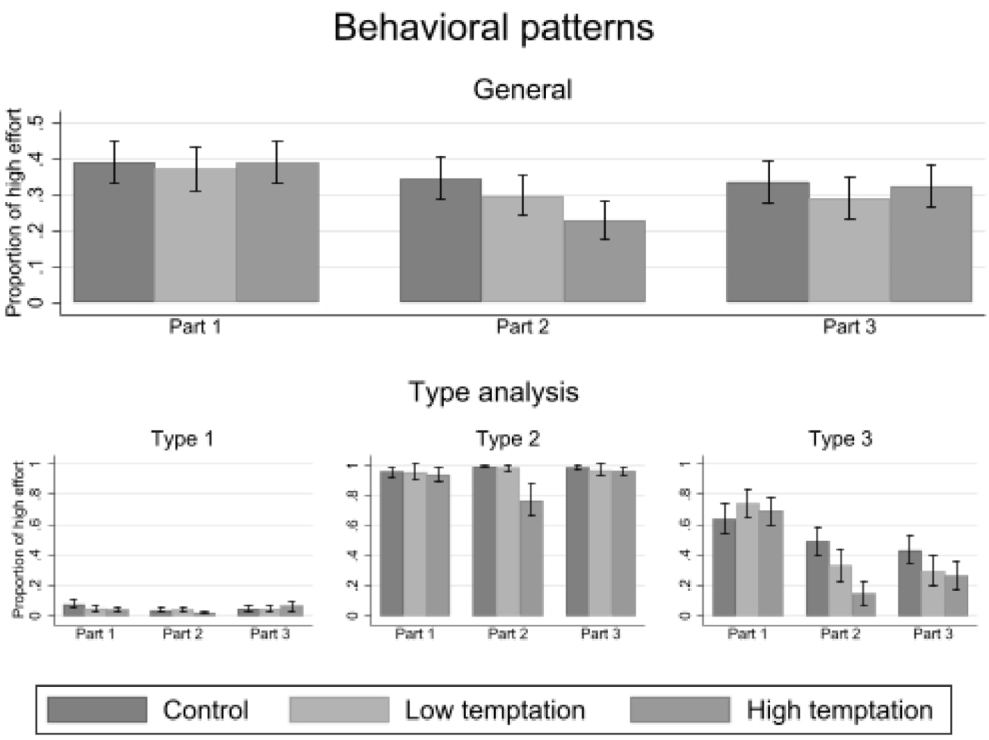


**Fig. 1.** Average proportion of large tables over time. The top panel displays the overall pattern, the bottom panel shows type-specific patterns. Error bars represent +/− one standard deviation.

Table 1 reports an OLS estimation in which we regress an indicator variable which is unity if a participant chooses a large table in a given period on dummy variables for treatments (Low, High), parts of the experiment (Part 2, Part 3) and the interaction effects of those variables. The p-values we report are based on the estimations from Table 1.

In Part 1 (identical conditions for all participants) the proportion of large tables is very similar in all conditions (C1: 0.39, L1: 0.37, H1: 0.39, $p_{C1,L1} = 0.656$, $p_{C1,H1} = 0.967$, $p_{L1,H1} = 0.685$).[1] In condition C in which participants face the same temptation throughout, the proportion of large tables drops in part 2 (C2: 0.35, $p_{C1,C2} = 0.007$) and then remains roughly constant in part 3 (C3: 0.34, $p_{C2,C3} = 0.544$, $p_{C1,C3} = 0.009$).[2] In condition L, there is a statistically significant decrease in the proportion of large tables in part 2 (L2: 0.30, $p_{L1,L2} < 0.001$). As hypothesized, this effect is not reversed in part 3, when incentives return to the initial level (L3: 0.29, $p_{L2,L3} = 0.628$, $p_{L1,L3} < 0.001$). Yet, a diff-in-diff test reveals that the decreases from part 1 to part 2 in conditions C and L are not significantly different ($p_{C2-C1,L2-L1} = 0.283$). Moreover, the proportions of large tables in Part 3 do not differ significantly between conditions C and L ($p_{C3,L3} = 0.258$).

**Table 1**
Mean immediate and delayed treatment effects of the choices of high efforts for the whole sample, and the three types identified by the Finite Mixture Model. The standard errors for the pooled regressions and the FMM are clustered at the individual level.

| | Pooled regressions | | Type analysis | | |
|---|---|---|---|---|---|
| | Model 1 | Model 2 | Type 1 | Type 2 | Type 3 |
| Constant | 0.392*** | 0.414*** | 0.071*** | 0.988*** | 0.615*** |
| | (0.030) | (0.030) | (0.015) | (0.016) | (0.026) |
| Low*Part2 | −0.029 | −0.029 | 0.012 | 0.04 | −0.242*** |
| | (0.026) | (0.026) | (0.020) | (0.025) | (0.068) |
| High*Part2 | −0.114*** | −0.114*** | 0.011 | −0.157** | −0.408*** |
| | (0.030) | (0.030) | (0.020) | (0.057) | (0.056) |
| Low*Part3 | −0.027 | −0.027 | 0.024 | 0.034 | −0.267*** |
| | (0.030) | (0.030) | (0.025) | (0.033) | (0.068) |
| High*Part3 | −0.01 | −0.01 | 0.062* | 0.048† | −0.284*** |
| | (0.030) | (0.030) | (0.028) | (0.028) | (0.067) |
| Low | −0.019 | −0.019 | | 0.004 | |
| | (0.043) | (0.043) | | (0.019) | |
| High | −0.002 | −0.002 | | −0.015 | |
| | (0.042) | (0.042) | | (0.019) | |
| Part2 | −0.046*** | −0.005 | | −0.046** | |
| | (0.017) | (0.019) | | (0.017) | |
| Part3 | −0.055*** | 0.027 | | −0.054** | |
| | (0.021) | (0.027) | | (0.021) | |
| Round | | −0.004*** | | | |
| | | (0.001) | | | |
| $N$ | 18450 | 18450 | | | |
| $R^2$ | 0.011 | 0.012 | | | |
| Sigma | | | | 0.281 | |
| | | | | (0.007) | |

***$p < 0.001$.
**$p < 0.01$.
*$p < 0.05$.
†$p < 0.1$.

[1] We use the following notation for reporting p-values: for comparisons of parts within or across conditions: $p_{Ti,Sj}$, where $T$ and S stand for treatments (C, L, or H) and i and j stand for the part (1, 2, or 3), for diff-in-diff comparisons: $p_{Ti-Tj,Si-Sj}$.

[2] Brañas Garza et al. (2013) use a dynamic model (with lagged-dependent variables as described in Arellano and Bond (1991) to establish evidence for the presence of moral self-regulation in a setting with repeated decisions in a stationary environment. We establish a link to this type of research by analyzing our data of the control condition with a similar model (in our control condition participants make a sequence of 30 decisions in a constant decision environment). We estimated a model with lags = 2 ($n$ = 5400, with 408 instruments, controlling only for the parts of the experiments). We do not find support for self-regulation over time, as neither the decisions in $t-1$ nor in $t-2$ are significant predictors of the decisions in $t$ ($t-1$: $\beta = 0.09$, $p = 0.257$; $t-2$: $\beta = 0.087$, $p = 0.148$). Note that the lagged dependent variables are valid instruments, as the Arellano–Bond test for autocorrelation in second-differenced errors was not rejected (first-order: $z = -4.792$, $p < 0.001$; second-order: $z = 0.43362$, $p = 0.665$). This analysis reinforces our conclusion that our data does not provide support for the self-regulation hypothesis.

**Table 2**
Proportions of Type 1, Type 2, and Type 3 individuals.

| | Type 1 | Type 2 | Type 3 |
|---|---|---|---|
| Proportion | 0.573 | 0.229 | 0.197 |
| SE | (0.02) | (0.017) | (0.016) |

In condition H we observe a more pronounced drop in the proportion of large tables in part 2 (H2: 0.23, $p_{H1,H2} < 0.001$). This effect is partially undone in part 3 when incentives are set back to their initial level (H3: 0.33, $p_{H2,H3} < 0.001$). The diff-in-diff analysis shows that both the decrease in part 2 and the increase in part 3 are significantly larger in condition H than in condition L ($p_{L2-L1,H2-H1} = 0.007$ and $p_{L3-L2,H3-H2} < 0.001$) and condition C ($p_{C2-C1,H2-H1} < 0.001$ and $p_{C3-C2,H3-H2} < 0.001$). However, contrary to our predictions the proportion of large tables in the last part does not go beyond its initial level. In fact, the proportion of large tables in Part 3 is even lower than in Part 1 ($p_{H1,H3} = 0.003$). There are also no significant differences to part 3 in condition C and L ($p_{L3,H3} = 0.399$, $p_{C3,H3} = 0.770$). These findings are incompatible with moral cleansing.

At the fully aggregated level our findings are only partially compatible with our hypotheses. However, an F test for individual fixed effects indicated that individuals vary greatly in terms of their initial effort provision as well as their reactions to the treatments ($p < 0.001$). Whereas individual differences in social interactions are widely recognized (e.g., Engelmann et al. (2019)), moral standards can be difficult to manipulate exogenously and very little is known about the role of such standards in self-regulation mechanisms. Therefore, we propose to operate a classification of individuals into distinct behavioral types. The next section uses finite mixture models to investigate whether the general pattern hides type-specific effects.

### 3.2. Type analysis

We analyze type-specific patterns using a finite mixture model (FMM) that builds on Bruhin et al. (2020). Assuming that types may differ in initial moral behavior and in reactions to treatments, the FMM cleanly identified three types[3] (see Table 2).

We summarize below the results of the type analysis. The reported p-values were calculated from the cluster-robust variance covariance matrix resulting from the FMM algorithm.

***Type 1.*** Type 1 individuals essentially display purely self-interested behavior. These individuals only solve very few large tables in Part 1 of the experiment (0.055, averaged across conditions C1, L1, and H1) and do not show any relevant reactions to our treatments.

***Type 2.*** Type 2 individuals initially solve almost exclusively large tables, displaying strong morality (0.95, averaged across conditions C1, L1, and H1). In Part 2 of the experiment, they only react in condition H (H2: 0.77, $p_{H1,H2} < 0.001$), while the proportion of large tables remains very high in the other two treatments (C2: 0.99, $p_{C1,C2} = 0.816$ and L2: 0.99, $p_{L1,L2} = 0.914$).[4] Once the temptation is removed, the proportion of large tables in condition H returns to its initial level (H3: = 0.97, $p_{H1,H3} = 0.750$).[5]

***Type 3.*** Type 3 individuals display an intermediate level of initial morality (0.69, averaged across conditions C1, L1, and H1). Type 3 can therefore react to temptations and self-regulate beyond their initial level of morality. Type 3 individuals significantly decrease the proportion of large tables in Part 2 of Condition L (L2: 0.33, $p_{L1,L2} < 0.001$).[6] However, as predicted, no self-regulation follows and the proportion of large tables in Part 3 remains below its initial level ($p_{L1,L3} < 0.001$) and below the corresponding level in Condition C ($p_{C3,L3} < 0.001$). In Condition H the proportion of large tables in Part 2 decreases significantly more than in Condition L (H2: 0.15, $p_{H1,H2} < 0.001$, and $p_{L1-L2,H1-H2} = 0.002$).[7] In Part 3 of Condition H the proportion of large tables increases marginally relative to Part 2 (H3: 0.26, $p_{H2,H3} = 0.076$), but it neither reaches its initial level ($p_{H1,H3} < 0.001$), nor the corresponding level in Condition C ($p_{C3,H3} < 0.001$). However, the increase in Part 3 of Condition H is larger than the corresponding changes in the other two treatments ($p_{L2-L3,H2-H3} = 0.084$, $p_{C2-C3,H2-H3} = 0.011$).

## 4. Conclusion

At the aggregated level, we observe that weak temptations lead to moderate, but persistent increases in immoral behavior. Strong temptations lead to larger increases in immoral behavior but fail to trigger moral cleansing as the effects are only partially undone. Our results provide evidence showing that moral cleansing has been overestimated, confirming the conclusions of the meta-analysis by Blanken et al. (2015) on moral self-regulation. A main reason for this overestimation lies in the fact that past research did not include a clean reference point for what constitutes "regular" behaviors. Without such an appropriate baseline, it is impossible to assess whether self-regulation compensates for past transgressions, or individuals simply return to their usual behavior after having transgressed. Moreover, our type-based analysis reveals two important insights. First, the individual baseline level of morality is even more important than anticipated, as it affects both the reactions to temptation (i.e., the deviations) and the efficiency of moral cleansing. Second, whereas we find some support for our conjecture that the salience of deviations from baseline morality affects the dynamics, we do not find evidence for moral self-regulation, even when accounting for individual differences.

## Acknowledgments

We are thankful to Justin Buffat, Ralph Hertwig, and Sirio Lonati for helpful comments. This project has received funding from the HEC Research Fund 2019.

## References

Abeler, J., Falk, A., Goette, L., Huffman, D., 2011. Reference points and effort provision. Amer. Econ. Rev. 101 (2), 470–492.
Arellano, M., Bond, S., 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Rev. Econom. Stud. 58 (2), 277–297.
Blanken, I., van de Ven, N., Zeelenberg, M., 2015. A meta-analytic review of moral licensing. Pers. Soc. Psychol. Bull. 41 (4), 540–558.

[3] We selected a model with three types over two types because its goodness of fit (using the Akaike information and the Bayesian information criteria) and quality of distinction between types (using normalized entropy and the integrated completed likelihood criteria) were higher. Moreover, we discarded models with four and five types for parsimony reasons.

[4] Diff-in-diff comparisons between condition H and the other conditions are also significant ($p_{L1-L2,H1-H2} = 0.009$ and $p_{C1-C2,H1-H2} < 0.001$).

[5] The proportions of large tables in Part 3 do not differ significantly between conditions H and C ($p_{C3,H3} = 0.179$).

[6] The change is significantly larger than in condition C ($p_{C1-C2,L1-L2} < 0.001$).

[7] The change is significantly different from condition C ($p_{C1-C2,H1-H2} < 0.001$).

Bruhin, A., Janizzi, K., Thoeni, C., 2020. Uncovering the heterogeneity behind cross-cultural variation in antisocial punishment. J. Econ. Behav. Organ. 180, 291–308.
Chapman, H.A., Anderson, A.K., 2013. Things rank and gross in nature: a review and synthesis of moral disgust. Psychol. Bull. 139 (2), 300.
Dolan, P., Galizzi, M.M., 2015. Like ripples on a pond: behavioral spillovers and their implications for research and policy. J. Econ. Psychol. 47, 1–16.
Effron, D.A., Cameron, J.S., Monin, B., 2009. Endorsing obama licenses favoring whites. J. Exp. Soc. Psychol. 45 (3), 590–593.
Engelmann, J.B., Schmid, B., De Dreu, C.K., Chumbley, J., Fehr, E., 2019. On the psychology and economics of antisocial personality. Proc. Natl. Acad. Sci. 116 (26), 12781–12786.
Fischbacher, U., Föllmi-Heusi, F., 2013. Lies in disguise—an experimental study on cheating. J. Eur. Econom. Assoc. 11 (3), 525–547.
Brañas Garza, P., Bucheli, M., Espinosa, M.P., García-Muñoz, T., 2013. Moral cleansing and moral licenses: experimental evidence. Econ. Philos. 29 (2), 199–212.
Gneezy, U., Imas, A., Madarász, K., 2014. Conscience accounting: Emotion dynamics and social behavior. Manage. Sci. 60 (11), 2645–2658.
Lee, S.W., Schwarz, N., 2021. Grounded procedures: A proximate mechanism for the psychology of cleansing and other physical actions. Behav. Brain Sci. 44.
Merritt, A.C., Effron, D.A., Monin, B., 2010. Moral self-licensing: When being good frees us to be bad. Soc. Pers. Psychol. Compass 4 (5), 344–357.
Sachdeva, S., Iliev, R., Medin, D.L., 2009. Sinning saints and saintly sinners: The paradox of moral self-regulation. Psychol. Sci. 20 (4), 523–528.
Zhong, C.-B., Liljenquist, K., 2006. Washing away your sins: Threatened morality and physical cleansing. Science 3135792, 1451–1452.